journal of applied psychology

KENNETH E. CLARK, Editor
University of Rochester

Volume 51, 1967

Published bimonthly by the American Psychological Association, Inc.
Prince and Lemon Sts., Lancaster, Pa. 17604 and 1200 Seventeenth St., N.W.
Washington, D. C. 20036

Second-class postage paid at Lancaster, Pa.

Contents of Volume 51

Journal of Applied Psychology

VOL. 51, NO. 1 FEBRUARY 1967

DEVELOPMENT OF A BEHAVIOR TAXONOMY FOR DESCRIBING HUMAN TASKS: A CORRELATIONAL-EXPERIMENTAL APPROACH [1]

EDWIN A. FLEISHMAN

American Institutes for Research, Washington, D. C.

The need for identifying a set of unifying dimensions underlying skilled behavior is discussed. The issues bear on problems of generalizing principles from laboratory to operational tasks and from one task to another. Combinations of experimental and correlational approaches appear to be required. The conceptual framework and research strategy utilized by the author in his research on perceptual-motor abilities is described and its relevance to taxonomy questions discussed. The integrative nature of the framework developed is illustrated by a wide variety of studies, in laboratory and operational situations, ranging from those of skill learning and retention to the effects of environmental factors on human performance, and in the standardization of laboratory tasks for performance assessment.

A fundamental prerequisite to the effective operation of any man-machine system is the acquisition, performance, and retention of skilled behaviors. For many years psychologists have studied human learning and performance, often in applied contexts, under numerous task and environmental conditions, and have accumulated vast quantities of data. And yet, as new systems are conceived for the exploration of space, for defense, for command and control, it appears that much of the accumulated data and experience of the past are largely inapplicable and that the problems of skill identification, training, and performance must be restudied almost from scratch. Why is this the case? Why such a waste of prior findings?

Superficially, each new system differs from other systems with respect to application, mission, and technology, and apparently in its task demands on its human operators. No two task analyses are ever quite the same. No two systems ever have identical job requirements. No training device or simulator ever quite seems to fit the requirements of any system except the one for which it was developed. Is it reasonable to conclude that the tasks of men in systems are so varied that there are no common dimensions with respect to the basic abilities required, the types of training needed for job proficiency, or the degradation of skilled behavior under given environmental conditions?

The present lack of a set of unifying dimensions underlying skilled behavior would appear to require one to answer the question in the affirmative. Why else have so many prominent psychologists (e.g., Melton, Fitts, Gagné, Miller) called so often for so long for a method for classifying human tasks—for a "task taxonomy"?

The problem is not only one of finding ways to generalize principles from one operational system to another. It also involves the generalization of findings from the laboratory

[1] Contribution to symposium "Task Taxonomy and its Implications for Military Requirements," at the American Psychological Association Convention, Chicago, September 1965. Preparation of this paper was facilitated by Contract No. DA-49-193-MD-2632 with the Army Surgeon General's Office, and by National Science Foundation Grant No. GB-1742.

to operational tasks. One reason why much of current learning research in the experimental laboratory appears so sterile to those who try to apply it to real-life training situations is the lack of concern by learning psychologists for the problem of task dimensions. This, of course, is often true for laboratory studies of the effects of environmental factors, the effects of motivational variables, the effects of drugs, etc., on human performance. It is not too long ago when a favorite distinction was between "motor" and "mental" tasks or between "cognitive" and "noncognitive" tasks. Such distinctions are clearly not very helpful in generalizing results to new situations, which involve a complex array of tasks and skills not adequately described by such all-inclusive terms. Tasks selected in laboratory research are not often based on any clear rationale about the class of task or skill represented. Most learning theory is devoid of any concern about task dimensions, and it is this deficiency which, many feel, makes it so difficult to apply these theories in the real world of tasks and people. What is needed is a learning and performance theory which ascribes task dimensions a central role.

At the other extreme, one finds task descriptions, often developed by job and systems analysts, which are highly detailed and highly specific. These are extremely useful and necessary for a variety of purposes, but their very specificity limits the kinds of generalizations which can be made to new classes of skills and tasks.

Categories which conceive of man-task interactions in terms of classes of functions certainly would seem to be steps in the right direction. Thus, Gagné (1964) tends to use categories like discrimination, identification, sequence learning, problem solving; and Robert Miller (1965) uses terms like scanning, identification of cues, interpretation, short-term and long-term memory, decision making, etc. Psychologists working with the Lockheed approach use dimensions like tracking, vigilance, arithmetic, and pattern comparison (Mangelsdorf, 1965).

These categories may turn out to be highly useful in both (*a*) organizing psychological data into categories of consistent principles, or (*b*) allowing more dependable predictions from laboratory to operations and from one operation to another. The demonstration, of course, still needs to be made. One needs to recognize, naturally, that these approaches are initially arm-chair, rational descriptive approaches. And there is nothing wrong with this, provided the necessary experimental predictive work is carried out to test the utility of these systems. However, I am skeptical that any small number of categories is going to be successful. For example, one of the most careful systematic approaches is that developed by Alluisi (1965) in his extension of the Lockheed approach. I am sure that this work represents a major step forward in task standardization and that we will learn much from this work. However, the basic battery will consist of from six to eight tests covering categories such as vigilance processes, memory functions, communication functions, intellectual functions, procedural functions, etc. Everything known about the correlations among human performances indicates a greater degree of specificity than this and considerable diversity of function within these categories. Since there are many types of human functions within each sub-area, one probably cannot generalize too far within each area. It is my feeling that we will just have to admit that human performance is complex and consists of many components. The problem is to simplify the description as far as we can, seeking all the while to find the limits and generality of the categories developed. And the selection of measures, diagnostic and representative of these categories, is an empirical rather than an arm-chair question.

The major point of this presentation is that there are empirical-experimental approaches to developing task taxonomies and that these represent alternative ways of going about it. It is my feeling that we already know quite a bit about task dimensions from experimental-correlational studies already completed, and that these allow us to be much more specific about task dimensions than do the more general categorical terms previously described. And I believe that combinations of experimental and correlational methods can develop a taxonomy of human performance which is applicable to a large

variety of tasks and situations. Furthermore, such an approach yields empirical indexes (e.g., factor loadings) for diagnostic measures of the categories developed. I would like to elaborate these issues.

Conceptual and Methodological Framework

First, I would like to define some concepts which have been developed. Second, I would like to present some illustrative data of past and current work, and finally point up some directions in which we might go.

I find it useful to distinguish between the concepts of ability and skill. As we use the term, *ability* refers to a more general trait of the individual which has been inferred from certain response consistencies (e.g., correlations) on certain kinds of tasks. These are fairly enduring traits which, in the adult, are more difficult to change. Many of these abilities are, of course, themselves a product of learning and develop at different rates, mainly during childhood and adolescence. Some abilities (e.g., color vision) depend more on genetic than learning factors, but most abilities depend on both to some degree. In any case, at a given stage of life, they represent traits or organismic factors which the individual brings with him when he begins to learn a new task. These abilities are related to performances in a variety of human tasks. For example, the fact that spatial visualization has been found related to performance on such diverse tasks as aerial navigation, blueprint reading, and dentistry, makes this ability somehow more basic.

The term *skill* refers to the level of proficiency on a specific task or limited group of tasks. As we use the term skill, it is task oriented. When we talk about proficiency in flying an airplane, in operating a turret lathe, or in playing basketball, we are talking about a specific skill. Thus, when we speak of acquiring the skill of operating a turret lathe, we mean that this person has acquired the sequence of responses required by this specific task. The assumption is that the skills involved in complex activities can be described in terms of the more basic abilities. For example, the level of performance a man can attain on a turret lathe may depend on his basic abilities of manual dexterity and motor coordination. However, these same basic abilities may be important to proficiency in other skills as well. Thus, manual dexterity is needed in assembling electrical components, and motor coordination is needed to fly an airplane.

Implicit in the previous analysis is the important relation between abilities and learning. Thus, individuals with high manual dexterity may more readily learn the specific skill of lathe operation. The mechanism of transfer of training probably operates here. Some abilities may transfer to the learning of a greater variety of specific tasks than others. In our culture, *verbal* abilities are more important in a greater variety of tasks than are some other types of abilities. The individual who has a great many highly developed basic abilities can become proficient at a great variety of specific tasks.

Elsewhere (Fleishman, 1964; Gagné & Fleishman, 1959) we have elaborated our analysis of the development of basic abilities. This included a discussion of their physiological bases, the role of learning, environmental and cultural factors, and evidence on the rate of ability development during the life span. With this much conceptualization in mind, we can say that in much of our previous work one objective has been to describe certain skills in terms of these more general ability requirements.

The original impetus for this program was a very applied problem. While I was with the Air Force Personnel and Training Research Center, one of our missions was to build better psychomotor tests for the prediction of pilot success. The wartime Air Force program had been highly successful in developing such tests. For example, the Complex Coordination Test had consistent validity for pilots. This seems not surprising, since the test seemed to be a "job sample" of aspects of the pilot's job. The pilot does manipulate stick and rudder controls. But there were many tests which had substantial validity but did not at all "resemble" the pilot's job. Cases in point are the Rotary Pursuit and Two Hand Coordination Tests. And there were other tests which seemed to resemble aspects of the pilot's job but had no validity. So it seemed to me that the first step was to

discover the sources of validity in these tests. What ability factors were there in common to psychomotor tests which were common to pilot performance (see, e.g., Fleishman, 1953, 1956; Fleishman & Hempel, 1956)?

Perhaps a not too extreme statement is that most of the categorization of human skills, which is empirically based, comes from correlational and factor-analysis studies. Many of these studies in the literature are ill designed or not designed at all. This does not rule out the fact that properly designed, systematic, programmatic, correlational research can yield highly useful data about general skill dimensions. We can think of such categories as representing empirically derived patterns of *response consistencies* to task requirements varied in systematic ways. In a sense this approach describes tasks in terms of the common abilities required to perform them. As an example, let us take the term "tracking," a frequent behavioral category employed by laboratory and systems psychologists alike. But we can all think of a wide variety of different tasks in which some kinds of tracking are involved. Can we assume that the behavioral category of tracking is useful in helping us generalize results from one such situation to another? Is there a general tracking ability? Are individuals who are good at compensatory tracking also the ones who are good at pursuit tracking? Do people who are good at positional tracking also do well with velocity or acceleration controls? What happens to the correlations between performances as a function of such variations? It is to these kinds of questions that our program was directed.

Some Previous Research

In subsequent years we have conducted a whole series of interlocking, experimental, factor-analytic studies, attempting to isolate and identify the common variance in a wide range of psychomotor performances. Essentially this is laboratory research in which tasks are specifically designed or selected to test certain hypotheses about the organization of abilities in a certain range of tasks (see, e.g., Fleishman, 1954). Subsequent studies tend to introduce task variations aimed at sharpening or limiting our ability-factor definitions. The purpose is to define the fewest independent ability categories which might be most useful and meaningful in describing performance in the widest variety of tasks.

Our studies generally start with some gross area of human performance. Thus, we have conducted studies of fine manipulative performances (Fleishman & Ellison, 1962; Fleishman & Hempel, 1954a), gross physical proficiency (Fleishman, 1963, 1964; Hempel & Fleishman, 1955), positioning movements and static reactions (Fleishman, 1958a), and movement reactions (Fleishman, 1958b; Fleishman & Hempel, 1956).

Thus far, we have investigated more than 200 different tasks administered to thousands of subjects (*S*s) in a series of interlocking studies. From the patterns of correlations obtained, we have been able to account for performance on this wide range of tasks in terms of a relatively small number of abilities. In subsequent studies our definitions of these abilities and their distinctions from one another are becoming more clearly delineated. Furthermore, it is now possible to specify the tasks which should provide the best measure of each of the abilities identified.

There are about 11 psychomotor factors and 9 factors in the area of physical proficiency which consistently appear to account for the common variance in such tasks. Their labels are Control Precision, Multilimb Coordination, Response Orientation, Reaction Time, Speed of Arm Movement, Rate Control, Manual Dexterity, Arm-Hand Steadiness, Wrist-Finger Speed, Aiming. In the physical proficiency area, a program was recently completed and described in book form (Fleishman, 1964); the factors have names like Extent Flexibility, Dynamic Flexibility, Static Strength, Dynamic Strength, Explosive Strength, Trunk Strength, Gross Body Coordination, Gross Body Equilibrium, and Stamina.

Of course, there are detailed descriptions of the operations involved in each category (see Fleishman, 1960b, 1962, 1964, 1966, 1967); some of them are more general in scope than others. But it is important to know, for example, that it is not useful to talk about strength as a dimension, but that, in terms of what tasks the same people can

do well, it is more useful to talk in terms of at least four general strength categories which may be differentially involved in a variety of physical tasks.

Perhaps it might be useful to provide some examples of how one examines the generality of an ability category and how one defines its limits. The definition of the Rate Control factor may provide an example. In early studies it was found that this factor was common to compensatory as well as following pursuit tasks. To test its generality, tasks were developed to emphasize rate control, which were not conventional tracking tasks (e.g., controlling a ball rolling through a series of alleyways). The factor was found to extend to such tasks. Later studies attempted to discover if emphasis on this ability is in judging the rate of the stimulus as distinguished from ability to respond at the appropriate rate. A task was developed involving only button pressing in response to judgments of moving stimuli. Performance on this task did *not* correlate with other rate control tasks. Finally, several motion picture tasks were adapted in which *S* was required to extrapolate the course of a plane moving across a screen. The only response required was on an IBM answer sheet. These tasks did not relate to the core of tasks previously found to measure "rate control." Thus, our definition of this ability was expanded to include measures beyond pursuit tasks, but restricted to tasks requiring the timing of a muscular adjustment to the stimulus change.

A similar history can be sketched for each

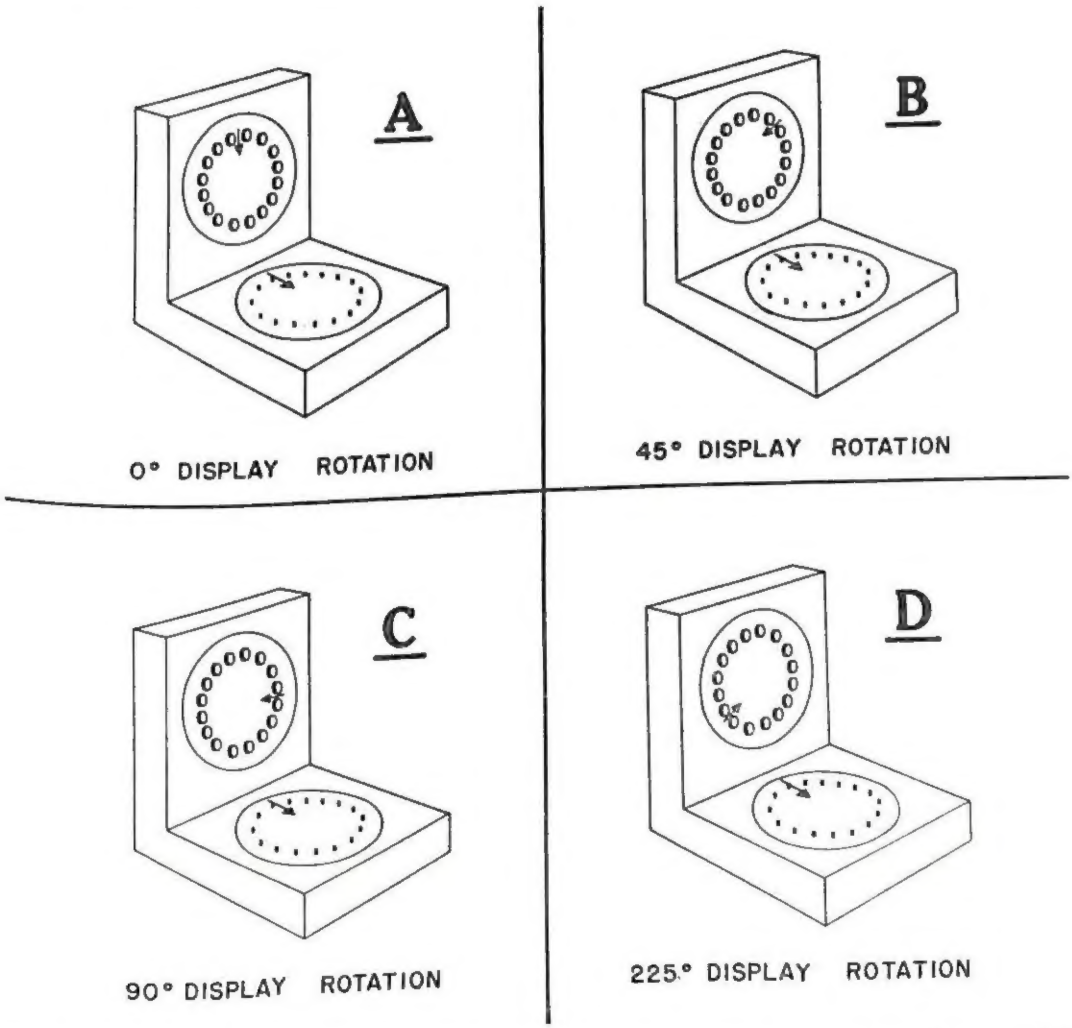


Fig. 1. The Response Orientation Device, showing four conditions of display rotation (drawing has omitted certain details not relevant). (After Fleishman, 1957.)

ability variable identified. Thus, we know that *S* must have a feedback indicator of how well he is coordinating before the Multilimb Coordination factor is measured; we know that by complicating a simple reaction-time apparatus, by providing additional choice reactions, we measure a separate factor (Response Orientation), but that varying the stimulus modality in a simple reaction-time device does not result in measurement of a separate factor.

Some later studies using experimental-correlational approaches provide encouraging results which indicate that it is possible to build up a body of principles through systematic studies of ability-task interaction in the laboratory. The approach is to develop tasks which can be varied along specified physical dimensions, to administer these tasks, systematically varied along these dimensions, to groups of *S*s who also receive a series of "reference" tasks, known to sample certain more generalized abilities (e.g., "Spatial Orientation," "Control Precision," certain "Cognitive Abilities"). Correlations between these reference tasks and scores on variations of the criterion task specify the ability requirements (and changes in these requirements) as a function of task variations. Thus far we have studied tasks varied along the following dimensions: degree of rotation of display panels relative to response panels; the predictability or nonpredictability of target course or response requirements; the extent to which the task allows *S* to assess the degree of coordination of multiple limb responses; the degree of stimulus-response compatibility in display-control relationships; whether there is a constant "set" or changing "set" from one stimulus presentation to the next; whether or not certain kinds of additional response requirements are imposed in a visual discrimination reaction task; whether or not certain kinds of feedback are provided. Hopefully, once such principles are established, it should be possible to look at new tasks, operational or otherwise, and specify the ability requirements.

TABLE 1

FACTOR LOADINGS OF RESPONSE MEASURES FOR DIFFERENT CONDITIONS OF DISPLAY ROTATION

Display rotation		Factor		
		Perceptual speed	Response orientation	Spatial orientation
0°	(↑)	.47	—	
45°	(↗)	.40	—	—
90°	(→)	—	—	.34
135°	(↘)	—	.37	.69
180°	(↓)	—	.40	.48
225°	(↙)	—	.30	.40
270°	(←)	—	—	.35
315°	(↖)	.36	—	.30

Note.—Loadings below .30 omitted. Adapted from E. A. Fleishman, 1957.

Perhaps I can illustrate with one study in which we were able to show systematic changes in the abilities required to perform a task as the display-control relations in the task were systematically varied. The *S* was required to press a button within a circular arrangement of buttons on a response panel, in response to particular lights which appeared in a circular arrangement of lights on the display panel (see Figure 1). The display panel was rotated from 0° to 45°, 90°, 135°, 180°, 225°, and 270°. The same *S*s performed under these task conditions and also performed on a series of spatial, perceptual, and psychomotor reference tasks. The results showed systematic changes in factors sampled by the criterion task as a function of degree of rotation of the display panel. Progressive rotation shifted the requirements from "Perceptual Speed" to two other factors "Response Orientation" and "Spatial Orientation" (see Table 1). Thus, individual differences along known dimensions are used to explore the relations between tasks and the characteristics of people who could perform the tasks most effectively. Of course, these are problems faced every day by personnel, training, engineering, and systems psychologists.

The results of other studies (e.g., the one in which we varied stimulus-response compatibility) showed that certain variations did make a difference but others did not. Such studies indicate that it should be possible to develop principles relating task dimensions to ability requirements, using laboratory experimental-correlational studies.

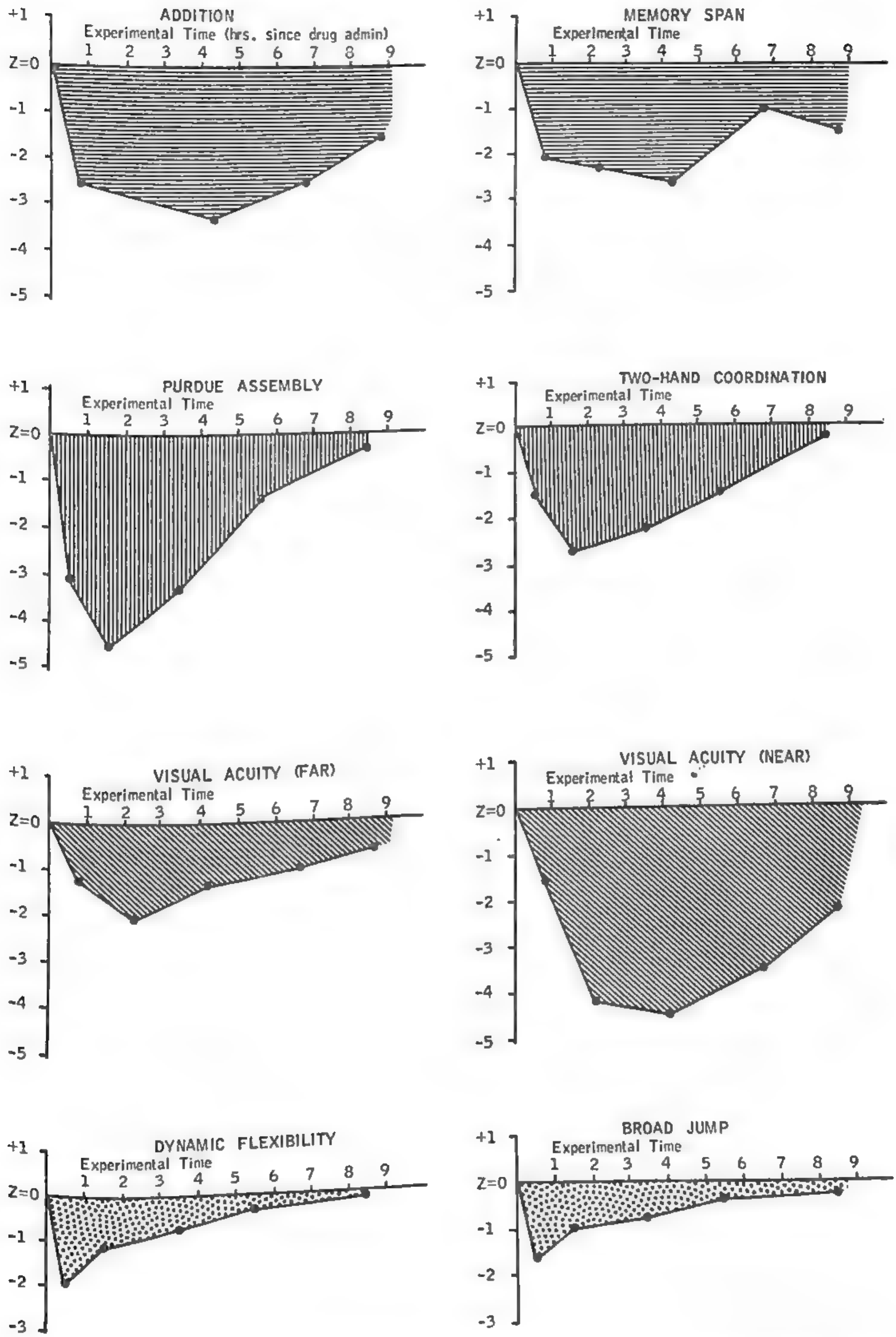


FIG. 2. Performance decrement on different ability tasks as a function of time since a given dosage of drug was administered. (After Elkin et al., 1965.)

Some Applications of the Taxonomy

I would like to mention a few areas in which the framework developed has proven useful in accounting for phenomena under investigation. For a number of years, I have been interested in the relations between reference ability measures, developed around our perceptual-motor taxonomy, and a variety of learning phenomena (see, e.g., Fleishman, 1957a, 1957b, 1960a; Fleishman & Hempel, 1954b, 1955; Fleishman & Rich, 1963). Our results have allowed us to show the differential role of these abilities at different stages of learning more complex tasks. These studies have as an additional goal the specification of abilities predictive of advanced levels of learning. The laboratory findings have been shown to hold with simulations of complex operational tasks such as piloting an air-intercept mission (Parker & Fleishman, 1960) and have been extended to real-life training situations involving skills other than perceptual-motor skills (Fleishman & Fruchter, 1960). We have recently summarized these findings elsewhere (Fleishman, 1966, 1967).

We have also used this paradigm to study part-whole task relationships (Fleishman, 1965; Fleishman & Fruchter, 1965), the prediction of associative interferences between two tasks, and the prediction of retention phenomena (Fleishman & Parker, 1962).

These categories have also proven useful in accounting for the interrelations among component proficiencies in fixed-wing (Fleishman & Ornstein, 1960) and helicopter piloting (Locke, Zavala, & Fleishman, 1965), and among driving proficiencies (Herbert, 1963). We have also been able to facilitate training in a simulated air-intercept tracking task using information developed on the role of these abilities at different stages of learning this complex task (Parker & Fleishman, 1961).

Are these categories useful as a means of task standardization and for generalizing research results to new tasks? We do not have any ready answer, but we are working on the problem. Let me at least mention a few studies in which tasks, representing these task dimensions, have been used. These studies have included research on the effects of stress—fear (Gorham & Orr, 1957), diet (Brozek, Fleishman, Harris, Lassman, & Vidal, 1955), and drugs (Baker, Elkin, Van Cott, & Fleishman, 1966; Elkin, Fleishman, Van Cott, Horowitz, & Freedle, 1965), and an ongoing study on the effects of high altitude (anoxia). For use in the space environment, in collaboration with James Parker, a console has been developed to measure in a compact unit 15 different factors for possible application in the manned orbital laboratory program (Parker, Reilly, Dillon, Andrews, & Fleishman, 1965). But it is in our drug work that we hope to make an approach to checking on the generalizations possible from standardized batteries of "basic ability" measures. We are observing the effects of a variety of drugs and dosages on measures of a variety of reference measures. These measures sample the perceptual, motor, sensory, and cognitive areas. We are getting differential effects, that is, some abilities within each area are more affected than others. Figure 2 gives a few illustrations of these events. The question is whether we get parallel results with other tasks representing the same factors. A further phase includes the development of complex tasks and the testing of these drug effects on such tasks. The question is whether our laboratory results, using component ability measures, could have predicted the drug effects on the complex tasks.

Finally, let me say that I am not at all sure about the ultimate utility of this kind of taxonomy. It may well be that the kind of taxonomy most useful to one set of applied problems (e.g., training) may be different from the one useful for another problem (e.g., selection, system design). But I am encouraged that the framework thus far developed has been useful in integrating some previously disparate data, that it has helped simplify some problems, and that it has led to some standardization of laboratory tasks used in a variety of research areas.

REFERENCES

ALLUISI, E. A., & THURMOND, J. B. Behavioral effects of infectious diseases: Annual progress report. *University of Louisville, Performance Research Laboratory Report,* 1965, No. PR-65-3.

BAKER, W. J., ELKIN, E. H., VAN COTT, H. P., & FLEISHMAN, E. A. *Effects of drugs on human performance: The development of analytical methods and tests of basic human abilities.* (Report AIR-E-25 3/66-TR2) Washington, D. C.: American Institutes for Research, 1966.

BROZEK, J., FLEISHMAN, E. A., HARRIS, S., LASSMAN, F. M., & VIDAL, J. H. Sensory functions and motor performance during maintenance of survival rations. *American Psychologist,* 1955, **10**, 502.

ELKIN, E. H., FLEISHMAN, E. A., VAN COTT, H. P., HOROWITZ, H., & FREEDLE, R. O. *Effects of drugs on human performance—Phase I: Research concepts, test development, and preliminary studies.* First Annual Summary Report. Washington, D. C.: American Institutes for Research, October 1965.

FLEISHMAN, E. A. Testing for psychomotor abilities by means of apparatus tests. *Psychological Bulletin,* 1953, **50**, 241–262.

FLEISHMAN, E. A. Dimensional analysis of psychomotor abilities. *Journal of Experimental Psychology,* 1954, **48**, 437–454.

FLEISHMAN, E. A. Psychomotor selection tests: Research and application in the United States Air Force. *Personnel Psychology,* 1956, **9**, 449–467.

FLEISHMAN, E. A. A comparative study of aptitude patterns in unskilled and skilled psychomotor performances. *Journal of Applied Psychology,* 1957, **41**, 263–272. (a)

FLEISHMAN, E. A. Factor structure in relation to task difficulty in psychomotor performance. *Educational and Psychological Measurement,* 1957, **17**, 522–532. (b)

FLEISHMAN, E. A. An analysis of positioning movements and static reactions. *Journal of Experimental Psychology,* 1958, **55**, 13–24. (a)

FLEISHMAN, E. A. Dimensional analysis of movement reactions. *Journal of Experimental Psychology,* 1958, **55**, 438–453. (b)

FLEISHMAN, E. A. Abilities at different stages of practice in Rotary Pursuit performance. *Journal of Experimental Psychology,* 1960, **60**, 162–171. (a)

FLEISHMAN, E. A. Psychomotor tests in drug research. In J. G. Miller & L. Uhr (Eds.), *Drugs and behavior.* New York: Wiley, 1960. Ch. 17. (b)

FLEISHMAN, E. A. The description and prediction of perceptual-motor skill learning. In R. Glaser (Ed.), *Training research and education.* University of Pittsburgh Press, 1962. Ch. 5.

FLEISHMAN, E. A. Factor analyses of physical fitness tests. *Educational and Psychological Measurement,* 1963, **23**, 647–661.

FLEISHMAN, E. A. *The structure and measurement of physical fitness.* Englewood Cliffs, N. J.: Prentice-Hall, 1964.

FLEISHMAN, E. A. The prediction of total task performance from prior practice on task components. *Human Factors,* 1965, **7**, 18–27.

FLEISHMAN, E. A. Human abilities and the acquisition of skill. In E. A. Bilodeau (Ed.), *Acquisition of skill.* New York: Academic Press, 1966. Ch. 3.

FLEISHMAN, E. A. Individual differences and motor learning. In R. M. Gagné (Ed.), *Learning and individual differences.* Columbus: Charles Merrill, 1967. Ch. 8.

FLEISHMAN, E. A., & ELLISON, G. D. A factor analysis of fine manipulative performance. *Journal of Applied Psychology,* 1962, **46**, 96–105.

FLEISHMAN, E. A., & FRUCHTER, B. Factor structure and predictability of successive stages of learning Morse Code. *Journal of Applied Psychology,* 1960, **44**, 96–101.

FLEISHMAN, E. A., & FRUCHTER, B. Component and total task relations at different stages of learning a complex tracking task. *Perceptual & Motor Skills,* 1965, **20**, 1305–1311.

FLEISHMAN, E. A., & HEMPEL, W. E. A factor analysis of dexterity tests. *Personnel Psychology,* 1954, **7**, 15–32. (a)

FLEISHMAN, E. A., & HEMPEL, W. E. Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika,* 1954, **18**, 239–252. (b)

FLEISHMAN, E. A., & HEMPEL, W. E. The relation between abilities and improvement with practice in a visual discrimination reaction task. *Journal of Experimental Psychology,* 1955, **49**, 301–312.

FLEISHMAN, E. A., & HEMPEL, W. E. Factorial analysis of complex psychomotor performance and related skills. *Journal of Applied Psychology,* 1956, **40**, 96–104.

FLEISHMAN, E. A., & ORNSTEIN, G. N. An analysis of pilot flying performance in terms of component abilities. *Journal of Applied Psychology,* 1960, **44**, 146–155.

FLEISHMAN, E. A., & PARKER, J. F. Factors in the retention and relearning of perceptual-motor skill. *Journal of Experimental Psychology,* 1962, **64**, 215–226.

FLEISHMAN, E. A., & RICH, S. Role of kinesthetic and spatial-visual abilities in perceptual-motor learning. *Journal of Experimental Psychology,* 1963, **66**, 6–11.

GAGNÉ, R. M. *Conditions of learning.* New York: Holt, Rinehart & Winston, 1964.

GAGNÉ, R. M., & FLEISHMAN, E. A. *Psychology and human performance: An introduction to psychology.* New York: Holt, Rinehart & Winston, 1959.

GORHAM, W. A., & ORR, D. B. *Research on behavior impairment due to stress.* Washington, D. C.: American Institutes for Research, 1957.

HEMPEL, W. E., & FLEISHMAN, E. A. A factor analysis of physical proficiency and manipulative skill. *Journal of Applied Psychology,* 1955, **39**, 12–16.

HERBERT, M. J. Analysis of a complex skill: Vehicle driving. *Human Factors,* 1963, **5**, 363–372.

LOCKE, E. A., ZAVALA, A., & FLEISHMAN, E. A. Studies of helicopter pilot performance: II. The analysis of task dimensions. *Human Factors,* 1965, **7**, 285–302.

MANGELSDORF, J. E. Empirical measurement: Rational approach. In H. P. Van Cott (Chm.), Task taxonomy and its implication for military requirements. Symposium presented at American Psychological Association, Chicago, September 1965.

MILLER, R. B. Task analysis and task taxonomy: Inventive approach. In H. P. Van Cott (Chm.), Task taxonomy and its implication for military requirements. Symposium presented at American Psychological Association, Chicago, September 1965.

PARKER, J. F., & FLEISHMAN, E. A. Ability factors and component performance measures as predictors of complex tracking behavior. *Psychological Monographs,* 1960, 74, No. 503.

PARKER, J. F., & FLEISHMAN, E. A. Use of analytical information concerning task requirements to increase the effectiveness of skill training. *Journal of Applied Psychology,* 1961, 45, 295–302.

PARKER, J. F., REILLY, R. E., DILLON, R. F., ANDREWS, T. G., & FLEISHMAN, E. A. Development of a battery of tests for measurement of primary perceptual-motor performance. Biotechnology, Inc. NASA Contract Report, 1965.

(Early publication received September 19, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 11-16

A STUDY OF REMOTE INDUSTRIAL TRAINING VIA COMPUTER-ASSISTED INSTRUCTION

H. A. SCHWARTZ AND H. S. LONG[1]

IBM Corporation, Poughkeepsie

During the latter half of 1965, several field engineers received their required training in new computer technology through remote computer-assisted instruction (CAI). Students at terminals located in 4 major cities communicated, through Tele-processing facilities, with a computer system located centrally. Students' examination scores, course completion times, and attitudes were compared with those of other students who received the material through self-study texts in use at the time. CAI students scored lower on 1 part of the examination, but completed the course in considerably less time than the self-study students. Attitude scores were somewhat equivocal. Students who had been exposed to both CAI and self-study texts indicated a strong preference for the former. When compared to a "regular classroom" type of presentation, however, the self-study students rated their method slightly higher than did the CAI students. CAI students' attitudes appear to be related to the availability of assistance when course material problems are encountered. Additional findings from locally trained CAI students are presented in support of this interpretation.

For several years research on the use of computers in education has been underway. In general, this research has been concerned with academic courses, environments, and students, at various scholastic levels.

Somewhat more recently, and paralleling this activity, efforts have been aimed at the application of computer-assisted instruction (CAI) to the problems of industrial training. Initial endeavors along this line were performed "in-house"—at a location that housed both the system and the student terminals. Thus, these students studied with the close supervision and assistance of the course authors and research personnel associated with the project.

In a later study (Schwartz & Haskell, 1966) the computer system was located in one building and the students in another. Thus, supervision and assistance were relaxed slightly. In cooperation with a manufacturing facility of a major electronics corporation, newly hired technicians received their required training in data-processing fundamentals through either CAI or the programmed-instruction (PI) text operationally in use at that time.

In a more recent phase of research, conducted during the latter half of 1965, student terminals were located even more remotely and in a number of widely dispersed locations. An extensive amount of subject matter was taught to field engineers in an operational, work-oriented environment. For comparative purposes, the students received their instruction through either the self-study texts in use at that time or through CAI. The remainder of this paper deals with the conduct and results of that study.

Method

Materials. The material employed in the study consisted of an introductory (preschool) course dealing with a new generation of computers. Satisfactory completion of this course was a prerequisite for further training at a distantly located education center. The operational form of this course consisted of a series of self-study textbooks written in a general programmed-text format. The topics covered included:

(*a*) Numbering Systems
(*b*) Programming Systems
(*c*) IBM System/360 Programming

[1] This study could not have been carried out without the major contributions made by R. K. Valley, who authored the CAI course employed in the study; Wayne Wengert, who provided invaluable additions and modifications to the computer-control program and dealt with program problems as they arose; and Leeland R. O'Neal, who authored some of the earliest CAI course material and who, along with R. K. Valley, supervised the general operation of the entire system.

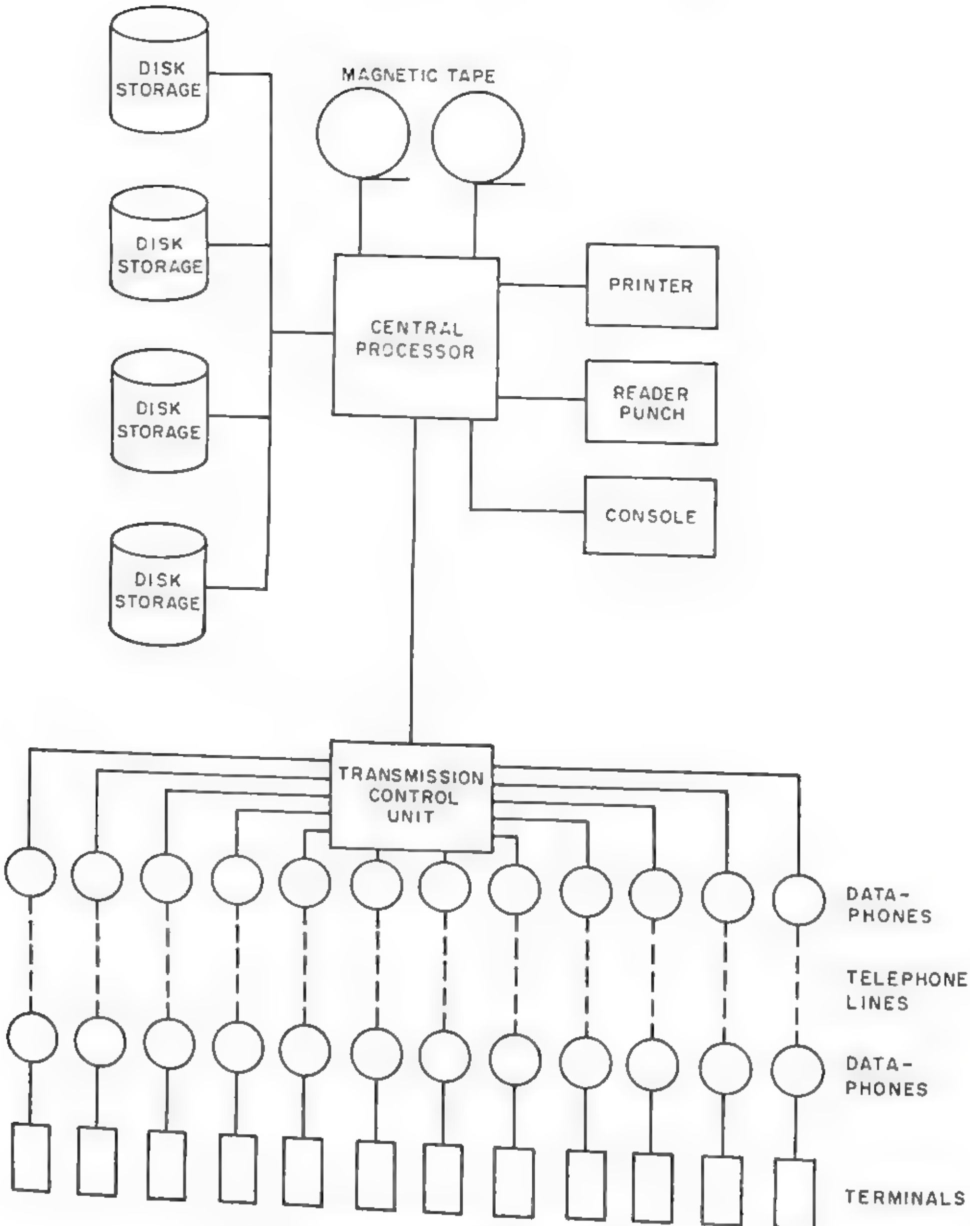


FIG. 1. Computer-assisted instruction system.

(d) Input/Output Programming
(e) SLT (Solid Logic Technology) Component Circuits and Packaging
(f) Core Storage

The CAI version of the course was designed to accomplish the same objectives as the self-study version.

CAI system. The CAI system employed in the study was an IBM 1440–1448 system, configured as shown in Figure 1. Although a number of student/author terminals are diagrammed in Figure 1, only four of these (one each in Los Angeles, San Francisco, Washington, D. C., and Philadelphia) were formally part of the study. The student terminals were of two types: one type consisted of a typewriter-like device on which the student could receive typewritten messages from the computer and could type his own messages to the computer. Visual material portions of the course text were contained in a separate book to which the student was referred, when necessary, by a typewritten message. The other type of terminal consisted of both the typewriter-like device and an associated stored-image visual display device. No book was necessary with the display device. The required visual and textual material was presented to the student under the control of the central computer. With both display and nondisplay terminals, the student responded to the system only via the terminal keyboard. For a more complete description of the CAI system, including examples of

student and author versions of course material, see Schwartz and Long (1966).

Subjects. During the predesignated pilot-study period of June 2–December 31, 1965, 40 field engineers participated in the study as students. At that time, each student was scheduled to participate in future advanced training at an education center. Students' background included experience with punch-card equipment, small and/or large computer systems. The distribution of students according to study location and method of instruction is shown in Table 1.

PROCEDURE

In keeping with its original intent—that of determining the feasibility of CAI as a means of accomplishing the required training without removing the student from his work location—the study was conducted under generally normal working conditions. In each location a monitor was designated to be responsible for operation of the terminal, the registration of students, maintenance of records, collection of data, and the establishment of study schedules.

As part of their normal educational sequence, students were assigned by their managers to complete the preschool course. Although a number of students might be taking the course at any given time, the original assignments were generally made on a one-student-at-a-time basis. When the assignment to the course had been made, the student was then further assigned to either the CAI or self-study technique of instruction by the study monitor. In the assignment an attempt was made to fully utilize the CAI terminal in each location. Thus for each student an attempt was first made to assign him to CAI. Where the terminal was not available at the time required by the student's schedule, he was assigned to the self-study group.

CAI students were ordinarily scheduled for study sessions of 2-hours duration, 5 days per week. In some cases, however, impending dates for education-center classes forced more extensive study periods.

Self-study students were less formally scheduled and supervised but were encouraged to spend at least 2 hours of each working day in study.

DATA

Three types of data were collected during the study. These were measures of (*a*) achievement, (*b*) study time, and (*c*) attitude.

Achievement. The achievement measures consisted of the scores attained on Parts II and III of an independently prepared, four-part final examination (Parts I and IV were not related to the study course material). The examination was multiple-choice in format and was administered to the students by the monitor. Part II consisted of 25 questions and Part III of 10. Each part was scored independently on a 100-point scale.

Time. Monitors maintained continuous records of on-job study time for the CAI students. Any off-job

TABLE 1

DISTRIBUTION OF STUDENTS BY LOCATION

Location	Number of students	
	Self-study	CAI
Los Angeles	5	4
Philadelphia	3	6
San Francisco	8	6
Washington, D. C.	0	8
Total	16	24

time spent in study (e.g., reading reference manuals) was reported on a questionnaire completed by the CAI students. The self-study students were supplied with a special form, on which they made daily recordings of their on-job and off-job study time, and which was returned to the monitor.

Attitude. At the end of the course, but prior to the final examination, each student completed an attitude questionnaire. This questionnaire solicited his comments on both the course material and the instructional method and also required him to compare his instructional method with other techniques.

HYPOTHESES

Based upon the previous research in CAI, the following hypotheses were advanced:

Hypothesis 1. Examination scores should be approximately equal for the two techniques. This would be expected on the basis that both the CAI and the self-study presentations were written to accomplish the same objectives. On the other hand, in view of the uncertainty inherent in examination scores of this type, it was stated that even if statistically significant difference should occur, no meaningful conclusions could be drawn unless such differences were quite large.

Hypothesis 2. The CAI students should complete the course in less time than the self-study students. This was based on the superior capability of CAI, through pretesting and system-controlled branching, to adapt to individual differences in ability and experience.

Hypothesis 3. The attitude of the CAI students toward CAI should be superior to the attitude of the self-study students toward self-study. This hypothesis was based on the ability of CAI to present a more individualized sequence of material and to provide hints when the student answers incorrectly.

RESULTS AND DISCUSSION

Achievement. The results shown in Table 2 tend to refute the original hypothesis that the two groups should score approximately the same on the examination. Nonparametric statistical tests, applied to these data, in-

TABLE 2

Mean Scores on Comprehensive Examination

Group	Exam section II	Exam section III
CAI	79.0	87.1
Self-study	87.4	92.9

dicate that the self-study group scored significantly higher on Section II of the examination (Mann-Whitney $U = 75$, $z = 2.42$, p (two-tailed) $< .02$). The same pattern appears on Examination Section III, but the difference does not attain statistical significance (p (two-tailed) $> .05$).

When the hypotheses of the study were set forth, it was noted that no differences were expected but that even if differences were to occur, the uncertainty involved in the interpretation of these examination scores would preclude the drawing of any firm conclusions unless the differences were quite large. In the case of Examination Section II, the actual difference obtained (8.4 points) is equivalent to approximately two multiple-choice questions on the test. In Section III, the difference (5.8 points) amounts to less than one multiple-choice question. Thus, while the obtained differences reached or tended toward statistical significance, these differences, presently, are of unknown practical significance.

Study Time Required

Table 3 shows the mean number of on-job, off-job, and total hours required to complete the course.

It may be seen that the two groups were approximately the same in terms of on-job time (62.5 and 64.3). However, the self-study

TABLE 3

Mean Hours Required to Complete Course

Group	*M* hours to completion: On job	Off job	Total
CAI	62.5	2.9	65.4
Self-study	64.3	12.3	76.6

students required significantly more off-job time (Mann-Whitney $U = 36.5$, $z = 9.9$, p (two-tailed) $< .0006$). This difference is reflected in the total time required to complete the course, which shows a savings of 11.2 hours, or approximately 15% in favor of the CAI group (Mann-Whitney $U = 101.50$, $z = 2.12$, p (two-tailed) $= .03$). These results are consistent with the original hypothesis that the CAI group should require less study time than the self-study group.

Attitude. In one portion of the attitude questionnaire, the students were asked to compare their method of instruction with others with which they were familiar. They did so by checking the most appropriate of five statements following each of the four questionnaire items. According to the statements they checked, their responses were assigned integer scale values ranging from 1 (indicating negative feelings toward their method) to 5 (indicating positive feelings toward their method). These results appear in Table 4.

It may be seen in Table 4 that in the comparisons with the regular classroom, the self-study method appeared to be rated slightly higher than the CAI method on each of the four questionnaire items. Only for Item III, however, was the difference between the two groups (3.3 versus 2.2) statistically significant (Median test, $\chi^2 = 5.75$, $df = 1$, p (two-tailed) $< .02$). In addition, note that in comparison with the regular classroom, the mean scale values for CAI were less than 3.0 for each of the four questionnaire items. Thus, it appears that the CAI students had somewhat negative feelings toward that method of instruction. This finding appears related to others to be discussed later in this report.

Inasmuch as the self-study students had not been exposed to CAI, no direct comparison of these two techniques could be made by them. On the other hand, the CAI students had experience with both techniques and consequently were able to rate one against the other. These results appear in the right-hand column of Table 4. Another nonparametric statistic (Kolmogorov-Smirnov one-sample test) was employed to determine whether these figures differ from what would be expected merely on the basis of "chance."

TABLE 4

MEAN SCALE VALUES ON ATTITUDE QUESTIONNAIRE ITEMS

Questionnaire item	Self-study compared with regular classroom	CAI compared with regular classroom	CAI compared with self-study
I. In your opinion, how well were you taught the material covered?	2.8	2.1	3.8
II. In your opinion, how difficult is it to learn through (CAI) (Self-Study)?	2.9	2.6	3.5
III. Which method of teaching do you like best?	3.3	2.2	4.2
IV. If you had your choice, which method would you use in future courses?	3.0	2.3	4.4

Note.—Scale values range from 1 to 5. A value of 1 indicates that students have negative feelings toward their own method of instruction. A value of 5 indicates positive feelings toward their own method of instruction. A value of 3 indicates indifference between their own and the comparison methods.

The results indicated that the responses to all but Questionnaire Item II were statistically significant ($p < .05$). Thus, it appears that students familiar with both techniques consider CAI superior to self-study.

Since this study represented the students' first use of this medium of instruction, it is, of course, possible that these attitudinal findings merely reflect the "novelty" of CAI, rather than any intrinsic instructional or motivational attributes it may possess. In view of the course length (65.4 hours), however, it is likely that any initial novelty effect would have dissipated prior to the completion of the material.

In the attitude questionnaire, the students also received a list of various instructional techniques and were requested to indicate the proportion of each technique that they would like to see employed in future theory-type courses they might take. The results of this survey are shown in Table 5. It may be seen that both the CAI and self-study groups indicated a clear-cut preference for their respective techniques to constitute the primary instructional method in future courses. The difference between these preferences (51.9 versus 62.7) was not, however, statistically significant.

The pattern apparent in Table 4 reappears in Table 5. In terms of the proportion preferred in future courses, the self-study techniques were rated as quite desirable by the self-study students (62.7%), but as rather unattractive in the eyes of the CAI students (4.8%).

It was previously noted that, in general, the CAI students tended to rate CAI somewhat lower than regular classroom instruction (see Table 4). Table 5, however, shows that these same students indicated approximately a 2:1 preference for CAI over regular classroom in future courses (51.9% versus 25.0%). One student, for example, gave CAI the lowest possible rating when comparing it with regular classroom. Yet he indicated that he would prefer to see CAI employed 90% of the time in future courses. Attitude-questionnaire comments and personal interviews conducted with CAI students provide a plausible explanation for this apparent paradox. By far, the students' major complaint was the unavailability of an instructor when they re-

TABLE 5

PREFERENCE FOR COMPOSITION OF INSTRUCTIONAL TECHNIQUE TO BE EMPLOYED IN FUTURE COURSES

Method	CAI group	Self-study group
Lecture by live instructor	25.0	18.0
Lecture presented by film or television	2.9	6.0
Self-study material (programmed texts, etc.)	4.8*	62.7*
Individual instructor to work with you	5.5	2.7
Discussion class (guided by instructor)	10.0	10.7
Computer-Assisted Instruction	51.9	—[a]

[a] Since the self-study students had no exposure to CAI, this alternative was not presented on their questionnaires.

* Significant difference between CAI and self-study (Median test, $\chi^2 = 18.9$, $df = 1$, $p < .0005$, two-tailed).

quired further explanation on some particular point in the course. The ability to type "help" and receive the correct answer (as is permitted by the system), but without any accompanying explanation, was unsatisfactory to them. In this respect, the text may even provide an advantage, since a student can more quickly skim the previously covered material and possibly find the solution to his problem. This may also account for the *suggestion* (the differences were not statistically significant) in Table 5 that the CAI students preferred more live lecture (25% versus 18%) and individual instruction (5.5% versus 2.7%) than did the self-study group.

Thus, it seems from their responses to the attitude questionnaire that the CAI students were indicating their desire for an instructor as something similar to a "fire escape"—they do not need him frequently or for long periods, but when they do, they feel it extremely important that he be immediately available. This interpretation is supported by data collected on five comparable students who completed the same course via CAI, but at the location of the central computer. (These data were not included in the previously presented results because these students were not formally a part of the study.) Under these conditions, immediate and expert help, in the person of the course authors, was continuously available. Because of the small sample size and the obvious possibility that some bias due to the transmitted enthusiasm of the CAI system personnel may have taken place, no statistical analysis of these data was attempted. However, in view of the interpretation previously advanced, it is interesting to note that these centrally located CAI students performed better than the remotely located CAI students in terms of both examination scores and study time.

When compared to both regular classroom and self-study, in terms of attitude, they also rated CAI considerably and consistently higher than did the remotely located CAI students.

Finally, and presumably because of the availability of assistance in their study location, the centrally located CAI students indicated a preference for more CAI and requested less individual tutorship and live lecture than did their counterparts in the more remote and less closely monitored CAI environments.

REFERENCES

SCHWARTZ, H. A., & HASKELL, R. J. A study of computer-assisted instruction in industrial training. *Journal of Applied Psychology*, 1966, 50, 360–363.

SCHWARTZ, H. A., & LONG, H. S. Instruction by computer: Motives and methods. *Datamation*, 1966, 12, No. 9.

(Early publication received August 26, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 17–19

HEREDITARY INFLUENCES ON VOCATIONAL PREFERENCES AS SHOWN BY SCORES OF TWINS ON THE MINNESOTA VOCATIONAL INTEREST INVENTORY[1]

STEVEN G. VANDENBERG AND RICHARD E. STAFFORD

University of Louisville School of Medicine

The Minnesota Vocational Interest Inventory (MVII) scores of 53 fraternal pairs of twins showed greater within-pair variances than did the scores of 71 identical pairs. For 9 of the scores and 4 of the area keys the F test showed a statistically significant increase in within-pair variance for the fraternal twins. This is interpreted as evidence for a hereditary component in interests in occupations requiring no scientific ability. The hereditary influence operates probably through personality variables.

The consensus of investigators in the area of vocational preferences seems to be that interests in occupations develop gradually as the child learns about himself and the world around him and that such interests only begin to take definite shape late in adolescence (Darley & Haganah, 1955; Roe, 1956; Super, 1957). However, the long-neglected finding of Carter (1932) that identical twins show greater similarity in occupational choices than do fraternal twins (as measured by the SVIB) would suggest that vocational preference is not entirely determined by environment. Carter's finding was later confirmed in replication by Vandenberg and Kelly (1964).

The idea that genes may help determine the occupation one prefers may seem rather startling at first thought. It is highly probable, however, that the aptitudes and personality traits disposing one to a particular vocational preference are under genetic influence, and thus the idea of hereditary influences on vocational preference becomes more acceptable.

In the studies of twins by Carter (1932) and Vandenberg and Kelly (1964) it was found that hereditary influence was most marked in the group of science-related occupations. To explore whether these hereditary influences could also be detected in vocational interests requiring a minimum of scientific ability, the Minnesota Vocational Interest Inventory (MVII), described by Clark (1961), was administered to 53 fraternal and 71 identical pairs of twins. The MVII measures interest in 20 nonprofessional occupations such as baker, painter, or truck driver. In addition, interests in 9 general areas are measured by homogeneous keys which group occupations requiring similar skills or interests such as office work, outdoor work, etc.

Several methods have been proposed for the assessment of the importance of hereditary components in a variable, such as a heritability index H (Holzinger, 1929), chi-square (Thurstone, Thurstone, & Strandskov, 1953), and analysis of variance (Dahlberg, 1926).

The latter method has been used in this study, following Clark (1956) and Kempthorne and Osborne (1961). The variance of the fraternal (DZ) twins is partitioned in the between-pairs and within-pairs components; similarly for the identical (MZ) twins.

First, one may wish to check that the between-pair variance exceeds the within-pair variance, that is, that the intraclass correlations are positive. Next, one obtains the ratio between the fraternal within-pair vari-

[1] The research reported here was supported by Grant HD 00843 and K3-MH-18382 of the National Institutes of Health. The serological tests on which the zygocity determination was based were performed by Jane Swanson of the Minneapolis War Memorial Blood Bank. The tests were scored by David C. Campbell and the calculations performed on the IBM 7090 at the University of Rochester through the help of K. E. Clark and Anne Naymik. The authors are grateful to the twins used in this study.

TABLE 1

COMPARISON OF FRATERNAL AND IDENTICAL WITHIN-PAIR VARIANCES ON 30 SCORES ON THE MINNESOTA VOCATIONAL INTEREST INVENTORY

Score	Girls 34 DZ—40 MZ	Boys 19 DZ—29 MZ	All cases 53 DZ—69 MZ
Baker	1.169	2.680**	1.796*
Carpenter	1.379	1.951*	1.583*
Electrician	1.415	1.395	1.331
Food service man	1.125	1.519	1.266
Hospital attendant	2.340**	2.458**	2.439**
Industrial education teacher	.941	1.031	.978
IBM operator	1.980	2.695**	2.247**
Milk wagon driver	2.204*	1.352	1.624*
Machinist	1.251	1.455	1.313
Painter	1.281	.665	.968
Plasterer	1.127	1.618	1.310
Plumber	.940	1.004	.931
Pressman	1.426	1.355	1.425
Printer	1.167	2.161*	1.513
Retail sales clerk	3.794**	1.994*	2.287**
Radio and TV repairman	1.045	1.852	1.368
Shipping and stock clerk	.849	1.418	1.071
Sheet metal worker	1.402	1.110	1.235
Truck driver	1.306	2.737**	1.820**
Truck mechanic	1.646	1.585	1.531*
Warehouseman	1.237	3.310**	1.935**
H_1—Machine repairs	3.826**	2.498*	2.503**
H_2—Medical hospital service	1.567	.991	1.353
H_3—Office work accounting	1.714*	.787	1.396
H_4—Radio, etc.	1.611	1.231	1.293
H_5—Food preparation and menu planning	1.107	1.188	1.139
H_6—Carpentry and furniture making	1.124	.924	1.003
H_7—Verbal activity aesthetic	1.007	.659	.813
H_8—Clean hands	1.412	.898	1.111
H_9—Athletics outdoor masculine	.766	1.090	.898

* $p < .05$.
** $p < .01$.

ance to the identical within-pair variance

$$F = \frac{\sigma^2 w\mathrm{DZ}}{\sigma^2 w\mathrm{MZ}}.$$

The identical within-pair variance is solely due to environmental influences, because identical twins, who arise from one fertilized egg, have the same genes, except for a rare somatic mutation here or there. The fraternal within-pair variance is due to similar environmental influences plus differences in genes, because fraternal twins result from two fertilized eggs. They will have, on the average, only 50% the same genes, just like other pairs of siblings not born at the same time. This ratio may be thought of roughly as between variance due to heredity plus environment and variance due to heredity alone. The statistical significance of the departure from 1.00 is evaluated as an F ratio.

Table 1 shows that significant F ratios were found more frequently for boys than for girls. With boys and girls combined, significant F ratios were found for nine occupational interests and four homogeneous vocational areas. Significant beyond the .01 level of probability were: baker, hospital attendant, IBM operator, retail sales clerk, truck driver, and warehouseman. Significant beyond the .05 level of probability were carpenter, milk wagon driver, and truck mechanic.

The F ratios for the homogeneous area scales reached the .01 level of significance for occupations involving machine repairs (Scale H_1), and the .05 level of significance for occupations involving general office work (H_3) and for those involving repair and operation of electrical and electronic devices (Scale H_4).

Conclusion

The results seem to indicate that hereditary influences on vocational interests are not necessarily limited to high-level abilities such as those required for scientific and professional occupations, but range over the entire occupational spectrum. It is most likely that the hereditary components influence one's aptitudes and personality traits, which in turn influence vocational preferences.

REFERENCES

CARTER, H. D. Twin similarities in occupational interests. *Journal of Educational Psychology*, 1932, 23, 641–655.

CLARK, K. E. *The vocational interests of non-professional men.* Minneapolis: University of Minnesota Press, 1961.

CLARK, P. J. The heritability of certain anthropometric characters as ascertained from measure-

ments of twins. *American Journal of Human Genetics*, 1956, 8, 49–54.

DAHLBERG, G. *Twin births and twins from a hereditary point of view*. Stockholm: Tidens Tryckeri, 1926.

DARLEY, J. G., & HAGANAH, T. *Vocational interest measurement*. Minneapolis: University of Minnesota Press, 1955.

HOLZINGER, K. J. The relative effect of nature and nurture influences on twin differences. *Journal of Educational Psychology*, 1929, 20, 241–248.

KEMPTHORNE, O., & OSBORNE, R. H. The interpretation of twin data. *American Journal of Human Genetics*, 1961, 13, 320–339.

ROE, A. *The psychology of occupations*. New York: Wiley, 1956.

SUPER, D. E. *The psychology of careers*. New York: Harper, 1957.

THURSTONE, T. G., THURSTONE, L. L., & STRANDSKOV, H. H. A psychological study of twins. I. Distribution of absolute twins differences for identical and fraternal twins. Report No. 4, Psychometric Laboratory, University of North Carolina, June 1953.

VANDENBERG, S. G., & KELLY, L. Hereditary components in vocational preferences. *Acta Geneticae Medicae et Gemellologiae*, 1964, 23, 266–277.

(Early publication received September 15, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 20–27

JOB SATISFACTION AND THE DESIRE FOR CHANGE[1]

EINAR HARDIN

School of Labor and Industrial Relations, Michigan State University

Questionnaire data on office employees' general readiness for change, satisfaction with existing amounts of 14 job aspects, and desire for job-aspect change are used in testing the hypothesis that a person's desire for specific changes is governed not only by the discrepancy between the attractiveness to him of existing and potential job characteristics but also by his assessment of the very process of change. The hypothesis is upheld by results of multiple-regression analyses of aggregate scores and of data for several individual job aspects.

The ability of rank-and-file employees and lower levels of management to delay or impede technical and organizational change has long been recognized by social scientists and by the managerial groups in which forces for change often originate. The resultant concept and phenomenon of resistance to change have been the focus of many articles in managerial literature (Schleh, 1960; Selekman, 1945; Zander, 1950) and of an increasing number of empirical and theoretical research studies (Coch & French, 1948; French, Israel, & Aas, 1960; French, Ross, Kirby, Nelson, & Smyth, 1958; Ginzberg & Reilley, 1957; Lawrence, 1954, 1958; Lippitt, Watson, & Westley, 1958; McMurray, 1947; Miller, 1960; Pagès, 1959; Ronken & Lawrence, 1952; Stewart, 1957; Vander Zaden, 1959).

In recent years, however, the possibility that employee attitudes toward change may extend beyond the range of hostility and indifference into the positive dimension of welcoming and seeking change has been discussed seriously by Lippitt et al. (1958) and by Jacobson (1959). The concept of readiness for change presented by the latter was studied empirically by Trumbo (1958, 1961), Jacobson, Trumbo, Cheek, and Nangle (1959), Faunce (1960), and Nangle (1961). The attitude measured by responses to a set of nine Likert-type items was found to have at most a weak negative relationship statistically to job satisfaction but to be related positively to favorability toward perceived specific changes (Trumbo, Nangle). Nonverbal behavior toward change, specifically participation in an employee suggestion plan, appeared unrelated to the person's own readiness for change (Hardin, 1964), but it was unclear whether suggestions were usually intended to generate changes in the work of suggesters or whether they sought instead to remove the need for further adjustments by causing changes in the work of others. At the same time, participants in the plan usually worked under supervisors whose readiness for change was intermediate, rather than high or low, and who may therefore perhaps have been permissive, albeit not eager, toward efforts to generate modifications in work methods or environment.

Logical analysis of the nine slightly modified items used in work by Faunce, Hardin, and Nangle suggests some heterogeneity. Two of the items (Nos. 1 and 5) seem to reflect the respondent's attitude toward a person generating changes in his own work, while the rest ostensibly cover his attitude toward having his work changed by others. The latter attitude, which might be regarded as passive readiness for change, is likely to dominate, by sheer number of items, the active readiness possibly reflected by the other two items. Furthermore, one item (No. 8) appears to express passive readiness for change in terms of the salary increase needed to compensate the respondent for a job shift, which unfortunately makes his objective or subjective economic circumstances a part of measured readiness. In contrast, six items measure passive readiness in more direct, affective terms. Nevertheless, the main content of the set of nine items seems to be preference between constancy and change in job content, with no explicit specification of the states of affairs existing before and after the supposed change.

[1] This research was supported by the School of Labor and Industrial Relations, Michigan State University.

Although the respondent might perhaps specify implicitly some of the characteristics of the status quo ante and post, the set appears to reflect an attitude toward a *process* of change, initiated by the respondent or his environment.

The present paper offers and tests a primitive theory of employee desire for specific changes. When choosing between a potential state of affairs, described by a vector of job characteristics, and his present state of affairs, described by a quantitatively different vector of job characteristics, the employee will consider two factors: (*a*) how well off (not purely financially) would he be in the potential state of affairs, once he got there, as compared with his well-being in his current state of affairs?; and (*b*) what attractions (challenges) and repulsions (dangers and nuisances) would be involved in getting from here to there? Desire for specific changes is heightened both by superiority of the potential over the initial state of affairs and by positive desirability of the very process of change. If the change process itself is viewed as undesirable, little or no desire for specific changes need emerge, even when the status quo is inferior to the potential state of affairs. If the change process is viewed as desirable, desire for specific changes may emerge, even when the status quo is equal or superior to the potential state of affairs.

The additional satisfaction obtainable by changing from the status quo to an alternative but preferred state of affairs may be enjoyed for a short while or for a long time, depending among other things on the person's work-life expectancy. Being in the nature of a rate of flow, the additional satisfaction should be weighted by the expected permanence of the alternative state of affairs. Because any one distant year may be presumed to matter less than any one year lying in the near future, the weights should increase with permanence, but less than in strict proportion. If the weights are left implicit, the level of obtainable additional satisfaction will matter more, relative to the desirability of the change process, among those with long rather than short work-life expectancy.

This theory may be restated in testable form, if one makes the following assumptions: (*a*) the readiness-for-change score of the Jacobson group reflects the respondent's assessment of the desirability of the change process itself; (*b*) a conventional job-satisfaction score dealing with specific job aspects reflects the desirability of present compared with potential job aspects; and (*c*) the frequency with which the respondent expresses a desire for change, in the direction of his choice, in the specific job aspects covered by the job-satisfaction score reflects the extent to which he wants specific changes in his job.

Statistically the theory then has a number of testable implications:

1. The desire for specific changes is positively related to general readiness for change and negatively related to job satisfaction.
2. The desire for specific changes regresses more strongly (the partial regression coefficient being larger in absolute terms) upon job satisfaction than upon general readiness for change among those with long work-life expectancy than is the case among those with short remaining work life.
3. If the range of attitudes in a given sample of respondents is sufficiently large, one will find some respondents who desire change in job aspects, although they are completely satisfied with them, and some who desire no change in job aspects, although they are strongly dissatisfied with them.
4. Readiness for change, representing the subjective magnitude of adjustment required by the process of change, is a more important factor in the desire for change in some job aspects than in others. Desire for specific change regresses more strongly upon it when a change in the amount of a job aspect requires much adjustment by the individual, such as might be the case with a change in the amount of skill, than when a job-aspect change requires little adjustment, as might be true for a routine change in pay.

The rest of the paper describes a test of the first and second implications of the theory, reports findings bearing on the third and fourth implications, and suggests some improvements in research design.

RESEARCH SITE AND DATA

The material for this study was obtained from the first of two questionnaire surveys that had been conducted by the author and associates in the School of Labor and Industrial Relations as part of

an earlier longitudinal study of employee response to technological change. The surveys took place in a medium-sized insurance company which in the intervening period installed an IBM 650 electronic computer for data processing. The questionnaires, which were identified, were administered by the staff of the School. More than 90% of eligible employees, including first-line and some higher supervisors, participated in each survey. Further details on the nature and administration of the questionnaires, the participation rates, and the available findings on employee response to technological change can be found elsewhere (Faunce, 1960; Hardin, 1960; Nangle, 1961).

The questionnaires completed by the 246 employees who participated in both surveys constituted the basic data pool. Questionnaires lacking answers to one or more of 37 items dealing with desire for change, job satisfaction, and readiness for change were eliminated. This strict rule reduced the study sample to 199 persons.

Information on desire for change in 14 job aspects and on satisfaction with the same aspects was obtained from a checklist contained in the first survey. The aspects were worded as shown in Table 3. The employees were asked to indicate whether they were "completely satisfied," "very satisfied," "quite satisfied," "somewhat satisfied," or "not satisfied" with each aspect. They were also asked to indicate whether they wanted to see "an increase," "no change," or "a decrease" in that aspect. (The concepts of increases and decreases in a job aspect were presumably familiar to the respondents, because questions about perceived change in these aspects and about computer impact upon them had been asked at earlier stages in the questionnaire, and because illustrative examples were given in connection with each of the checklists in which the questions were asked.)

A *desire for change* in a job aspect was regarded as present when the respondent indicated he wanted an increase or a decrease in that aspect. Some of those who checked the "no change" category probably wanted to preserve the status quo, while others may have been indifferent toward possible changes. A desire-for-change score was computed for each person by counting the number of items for which he indicated he wanted a change.

TABLE 1

PRODUCT-MOMENT CORRELATIONS AMONG ITEMS

Item set	Size of correlation coefficient		
	Lowest	Median	Highest
Readiness for change	−0.06	0.20	0.46
Job satisfaction	0.19	0.51	0.72
Desire for change	0.00	0.26	0.55

Note.—This table is based on data for 199 respondents. The distributions of correlation coefficients are based on 36 pairs of items in the readiness-for-change scale and on 91 pairs in the other two scales.

TABLE 2

FREQUENCY DISTRIBUTIONS OF INTERCORRELATIONS AMONG READINESS-FOR-CHANGE ITEMS

Coefficient of correlation	Among items nos. 1,5,8	Among items nos. 2,3,4, 6,7,9	Between items nos. 1,5,8, and nos. 2,3,4,6,7,9
−.10 −.01	1		
+.00 .09	2		5
.10 .19		1	8
.20 .29		3	5
.30 .39		6	
.40 .49		5	
Total frequency	3	15	18

Note.—N = 199.

The *level of satisfaction* with an aspect was expressed numerically by assigning the arbitrary weights of 1, 2, 3, 4, and 5 to the five categories of response, from "completely satisfied" to "not satisfied," and a job-satisfaction score was computed for each person by summating the 14 weights corresponding to the response categories he checked and reversing the direction of scoring.

Readiness for change was measured by responses to slightly modified versions of the nine items used by Trumbo. Eight of these were statements to which the employee was asked to respond by stating that "I strongly agree," "I agree a little," "I neither agree nor disagree," "I disagree a little," and "I strongly disagree." The statements were as follows:

1. If I could do as I pleased, I would change the kind of work I do every few months.
2. One can never feel at ease on a job where the ways of doing things are always being changed.
3. The trouble with most jobs is that you just get used to doing things in one way and then they want you to do them differently.
4. I would prefer to stay with a job I know I can handle than to change to one where most things would be new to me.
5. The trouble with many people is that when they find a job they can do well they don't stick with it.
6. I like a job where I know that I will be doing my work about the same way from one week to the next.
7. When I get used to doing things in one way, it is disturbing to have to change to a new method.
8. It would take a sizable raise in pay to get me to accept a different job here.

The items were presented in the above order, but not in the form of a checklist. The ninth item was of the controlled-completion type, that is, The job that you would consider ideal for you would be one where the way you do your work: is always the same; changes very little; changes somewhat; changes quite a bit; changes a great deal. (These

items were separated by more than a dozen mimeographed pages from the checklists for job satisfaction and desire for specific changes.) A total score on readiness for change was computed by assigning values from 1 to 5 to the response categories ranked in ascending order of readiness and summating the weights.

The distributions of item correlations are summarized in Table 1. The reliability of the two 14-item scores appeared to have been fairly high. However, the readiness-for-change items appeared fairly heterogeneous. Further inspection of the correlation matrix for readiness-for-change items showed that the three items discussed earlier (Nos. 1, 5, and 8) were not very closely related to each other and to the other six items, as can be seen from Table 2. This finding suggested that one score should be computed for the set of six passive readiness-for-change items and that the other three items might be used as separate variables in the regression analysis.

RESULTS

The ordinary least-squares procedure was used for estimating the constants and partial regression coefficients of a linear multiple-regression equation which had the 14-item desire-for-change score as dependent variable (y), and which had as independent variables the 14-item job-satisfaction score (x_1), the 6-item passive readiness-for-change score (x_2), and the scores on the remaining individual items Numbers 1, 5, and 8, or x_3, x_4, and x_5. The resulting regression equation was:

$$\begin{array}{llllll} \hat{y} = 8.858 & - 0.150x_1 & + 0.271x_2 & - 0.312x_3 & - 0.346x_4 & + 0.104x_5 \\ & (0.021) & (0.046) & (0.162) & (0.178) & (0.163) \\ & p < .01 & p < .01 & p > .05 & p > .05 & p > .05 \end{array} \quad [1]$$

with $R^2 = 0.33$, corrected for degrees of freedom. The coefficient of multiple correlation was statistically highly significant, as were the first two partial regression coefficients.

As may be inferred from the standard errors, shown in parentheses, the desire for specific changes did not regress significantly on the three separate readiness-for-change items. Further analysis showed that there was also no significant regression on a score formed by adding these three items and that the value of the corrected R^2 remained virtually unchanged both when the 3-item score was substituted (0.32) and when all three items were eliminated from the regression equation (0.31). Since variables x_3 and x_4, as well as x_5, were only weakly, if at all, associated with the items aggregated into variable x_2, this finding indicated that high active readiness for change did not lead to a high desire for specific change in job aspects. Curiously enough, the regression weights for the two active-readiness items were negative, while the regression weight for the 6-item passive-readiness score was positive. Since the former two regression weights were not statistically significant, the negative signs were probably artifacts calling for no particular explanation.

The subsequent analysis involving readiness for change was confined to the score based on the six passive-readiness items. After variables x_3, x_4, and x_5 were dropped, a simplified regression equation was computed:

$$\begin{array}{llll} \hat{y} = 7.683 & - 0.146x_1 & + 0.245x_2 & \bar{R}^2 = 0.31 \\ & (0.020) & (0.043) & df = 196 \\ & p < .01 & p < .01 & \end{array} \quad [2]$$

Because job satisfaction (x_1) and readiness for change (x_2) were virtually uncorrelated ($r_{12} = -0.09$), the regression coefficients were not likely to have been affected importantly by multicollinearity. The partial correlation with desire for change

$$(r_{y1} = -0.45,\ r_{y2} = 0.38)$$

were both very significant. Statistically, therefore, desire for specific changes was associated inversely with job satisfaction and positively with passive readiness for change, so that the first implication of the theory was supported by the evidence.

The regression equation, while highly significant, left unexplained as much as 69% of the variation in desire-for-change scores. A variance analysis was made of the residuals (errors of estimate) for respondents grouped according to whether they were high, medium, or low on readiness for change and on job satisfaction, but it failed to show any statistically significant differences ($p > .20$). Whether the equation represented the data adequately at the logical extremes of the independent variables could not be directly tested, because few respondents had scores close to the boundary lines. When respondents

indicated being "not satisfied" with an individual job aspect, they invariably wanted a change in it. This fact suggested either nonlinearity or errors in regression estimates. Furthermore, the regression equation predicted that if persons could be found who were completely satisfied with all job aspects, and not at all ready for change, they would show a frequency of desire for change, averaging −1.3 aspects. This logically absurd prediction might also indicate nonlinearity at extreme values of independent variables, but the discrepancy from zero desire for change lay within the boundaries resulting from random errors in the regression coefficients. Since the respondents were classified into the three groups of high, medium, and low on each independent variable for the analysis of variance and since the F ratios were all below 1.0, it seemed reasonable for general purposes to disregard the weak evidence to the contrary and to regard the regression as linear in each variable and additive.

Two interesting implications of Regression Equation 2 should be noted. First, if a person is very unready for change ($x_2 = 6$), he will not desire change in all job aspects ($y = 7.7$), even if he is "not satisfied" with each of them ($x_1 = 14$). Similarly, if he is very ready for change ($x_2 = 30$), but is "completely satisfied" with each of the job aspects ($x_1 = 70$), he will nevertheless want to see change in some of them ($y = 4.3$). This tends to support the third main implication of our theory.

The questionnaire also contained the item, "How many more years do you expect to work?" The response categories were "less than 1 year," "1–2 years," "3–5 years," "6–10 years," "11 or more years," and "I don't know." The 199 persons were grouped into four categories of expected work life as shown in Table 3, and separate regressions were run to test the second implication, that job satisfaction increases in importance relative to readiness for change, as expected work life is increased. The results did not contradict the general proposition that desire for specific change regresses positively on job satisfaction and negatively on readiness for change, although some coefficients were too small to be significant. The second implication, on the contrary, was definitely not upheld by the results. While the differences among regression coefficients for job satisfaction were not statistically significant, the coefficients seemed to decrease with increased expected work life and did not rise in relation to the regression coefficients of readiness for change. This result suggests that the expected durations of the two states of affairs do not affect the desire for specific change. One other statistical result cautions against accepting this conclusion firmly.

While job satisfaction and readiness for change were approximately independent in the last three categories ($r = 0.06$, 0.12, and −0.00, respectively), they were rather highly correlated ($r = -0.50$) in the group having a remaining work-life expectancy of up to 2 years. This may have biased the relative magnitudes of the two regression coefficients in that group, but the direction of the bias is not obvious.

TABLE 3

EFFECTS OF WORK-LIFE EXPECTANCY UPON REGRESSION RESULTS

Statistic	Remaining work-life expectancy			
	0–2 years	3–10 years	11 or more years	Don't know
Sample size	38	32	52	77
R^2 corrected for df's	0.31	0.29	0.17	0.25
Job satisfaction				
Regression coefficient	−0.20	−0.13*	−0.15	−0.12
Standard error	0.06	0.05	0.04	0.03
Readiness for change				
Regression coefficient	0.07[a]	0.30	0.09[a]	0.21
Standard error	0.11	0.10	0.10	0.07

Note.—Correlation and regression coefficients are significant at the 1% level, unless otherwise stated.
[a] *ns* at the 5% level.
* $p < .05$.

The fourth implication of our theory was tested by analysis of each of the 14 job aspects. The desire for change in one job aspect would probably depend somewhat on the level of satisfaction with some other job aspects. This appeared particularly likely for pairs in which one aspect was pay, the prime compensation for jobs deficient in other regards. Nevertheless, the desire for change in any one aspect was not hypothesized to regress on satisfaction with other aspects, because the high intercorrelations among the 14 job-satisfaction items (see Table 2) would probably make the regression weights unduly sensitive to extreme observations. In brief, the desire for change of the i-th person in the j-th job aspect was hypothesized to be governed by the relation:

$$y_{ij} = a_j + b_{1j}\,x_{1ij} + b_{2j}\,x_{2i} + e_{ij} \qquad [3]$$

where

$$y_{ij} = \begin{cases} 1 \text{ if the i-th person wanted a change in aspect j} \\ 0 \text{ otherwise} \end{cases}$$

a_j = the constant term for aspect j

x_{1ij} = the i-th person's satisfaction with the j-th aspect

x_{2i} = the i-th person's score on the 6-item passive readiness-for-change scale common to all 14 equations

e_{ij} = the error or disturbance term for the i-th person on the j-th aspect.

The choice of both estimation method and significance tests would depend greatly on the relationships between the disturbance terms e_{ij} of the 14 equations. If the disturbance terms were uncorrelated, the ordinary single-equation least-squares method would be appropriate for estimating the regression coefficients and for conducting the required tests of significance. If they were correlated positively or negatively, it would seem preferable to use a two-stage estimation procedure which has been developed by Zellner (1962). According to Zellner this procedure "yields coefficient estimators at least asymptotically more efficient than single-equation least-squares estimators [p. 977]," particularly when the disturbance terms are highly correlated and the independent variables are not, and much of this gain is attained even in samples of 35 observations or less. However, it is slightly inferior to the single-equation approach, when the disturbance terms are uncorrelated and the observations are few.

The great similarity-in-choice situation which the respondent faced in reacting to different job aspects suggested that if some omitted variable was relevant in one job aspect, it would also be relevant in several other aspects and that individual variation in such a variable would therefore make a number of the residual correlations non-zero. For this reason the Zellner method was adopted. The calculations were performed on the Michigan State University CDC 3600 computer system using a program (ZEF) that had been developed by W. L. Ruble [2] from the Stroud and Zellner (1962) CDC 1604 version and that provided not only the efficient estimators and their appropriate standard errors but also calculated the requisite F statistics for tests of homogeneity of efficient estimators.

The residual correlation coefficients were found to range from -0.05 to $+0.50$, with a median value of 0.22. Judging by Zellner's (1963) Table 1 and the fairly large number of observations there was probably some gain, and little risk of serious loss, except for computing time, in using the Zellner approach. The efficient estimators, their corresponding standard errors, and the significance levels are shown in Table 4 for the two independent variables on each of the 14 job aspects.

The regressions of desire for change upon job satisfaction were negative and highly significant for all of the 14 job aspects. In contrast, only 8 of the regressions upon readiness for change were highly significant individually. Nevertheless, the signs of the regressions were nonnegative in all 14 cases. The main hypothesis of the paper was therefore strictly upheld in a majority of cases, and in no case was it contradicted by a significant regression having the wrong sign.

The table also suggests, however, that the regression relationship was not the same for all job aspects. Direct statistical tests of the hypothesis of homogeneity, namely, the equality of all 14 regression coefficients for each independent variable, confirmed this impression strongly: the test statistics were $F = 2.74$, and $F = 3.88$, respectively, as com-

[2] Department of Agricultural Economics, Michigan State University.

TABLE 4

REGRESSION ANALYSIS OF DESIRE FOR CHANGE: EFFICIENT ESTIMATORS FOR 14 JOB ASPECTS

Job aspect	Job satisfaction		Readiness for change	
	b^a	SE	b^a	SE
1. The amount of variety in my work.	−.224	.023	.026	.006
2. The amount of work required on my job.	−.189	.027	.024	.006
3. The degree of accuracy demanded by my job.	−.102	.030	.000[b]	.006
4. My control over the pace of my work.	−.226	.024	.012*	.006
5. The importance of my job for the company.	−.178	.025	.033	.006
6. The amount of supervision I get on my job.	−.176	.024	.004[b]	.005
7. The amount of skill needed on my job.	−.189	.028	.022	.006
8. The amount of responsibility demanded by my job.	−.204	.027	.014*	.006
9. The amount of planning I have to do on my job.	−.173	.026	.014	.006
10. The amount of judgment I have to use on my job.	−.208	.024	.022	.006
11. The degree to which my work is interesting.	−.257	.022	.020	.006
12. The amount of security I feel on my job.	−.238	.025	.007[b]	.006
13. My chances for promotion to a better job.	−.206	.019	.027	.005
14. The amount of pay I get on my job.	−.141	.019	.009*	.004

[a] All regression coefficients are significant at the 1% level, unless otherwise noted.
[b] *ns* at the 5% level.
* $p < .05$.

pared with $1.69 \leq F_{.05} \leq 1.76$ and $2.07 \leq F_{.01} \leq 2.20$ with $df = 13$ and 2744 (see Zellner, 1962, pp. 366–367). The null hypotheses of uniformity of regression coefficients among job aspects were therefore rejected.

If the desire-for-change components attributable to job satisfaction were definitionally identical with job satisfaction except for sign reversal, one would have expected no variation in regressions upon satisfaction from one aspect to the next. The statistically significant variations shown by the F test and Table 4 refuted the interpretation of definitional identity and suggested the presence of factors which influenced desire for change by interacting with job satisfaction. Among such factors might have been the degree of importance attached to being satisfied with a particular aspect and the subjective probability of seeing, or obtaining, a change in that job aspect.

The significant variations in the readiness-for-change component also indicated that the relationship to desire for change was not one of definitional identity. They might reflect differences in the length, hazards, and opportunities, or effort involved in the change process associated with movement from one job aspect to another. It seems natural that desired changes (typically increases) in the amount of variety and importance of the job and in the skill and judgment it required might call for substantial adjustments by the individual which would not be needed for a pay increase, which in this company was usually of a periodic, routine nature, or a change in the amount of supervision. But these explanations are ad hoc and call for systematic testing on the basis of additional variables.

DISCUSSION

The weakness of the (statistically very significant) relationship between job satisfaction and desire for specific changes, the variation in this relationship from one job aspect to another, and the presence of a significant regression of desire for change upon passive general readiness for change have an implication for research procedure: desire-for-change questions cannot safely be used to obtain simultaneously a measure of job satisfaction and an indication of the direction of change in job aspects needed to raise satisfaction. More fundamentally, general but passive readiness for change affects the frequency of desire for specific changes that may be expected at any given level of job satisfaction. This is true both as reflected by aggregate scores and for a number of individual job aspects. The re-

gression of desire for specific changes upon job satisfaction and general readiness for change varies significantly from one job aspect to another, probably in response to implicit but interacting variables which may include the importance of the job aspect to the individual, the subjective probability of experiencing a change in the aspect, and the nature and extent of the very process of change associated with an alteration in the amount of any particular job aspect.

Several improvements in research design suggest themselves: choice of more distinct and equidistant response categories for the job-satisfaction items, provision of several response levels for the individual desire-for-change items, and explicit introduction of additional explanatory variables. Replication of the study on the basis of such improved data may help throw further light on the main idea tested and supported in this analysis: the desire for specific changes in job aspects is governed not only by the discrepancy between the attractiveness of existing and potential job characteristics but also by the person's assessment of the very process of change.

REFERENCES

COCH, L., & FRENCH, J. R. P., JR. Overcoming resistance to change. *Human Relations,* 1948, **1,** 512–532. Reprinted in S. G. Hoslett (Ed.), *Human factors in management.* New York: Harper, 1951. Pp. 242–268.

FAUNCE, W. A. Social stratification and attitude toward change in job content. *Social Forces,* 1960, **39,** 141–148.

FRENCH, J. R. P., JR., ISRAEL, J., & AAS, D. An experiment on participation in a Norwegian factory. *Human Relations,* 1960, **31,** 3–19.

FRENCH, J. R. P., JR., ROSS, T. C., KIRBY, S., NELSON, J. R., & SMYTH, P. Employee participation in a program of industrial change. *Personnel,* 1958, **35**(3), 16–29.

GINZBERG, E., & REILLEY, E. W. *Effecting change in large organizations.* New York: Columbia University Press, 1957.

HARDIN, E. The reactions of employees to office automation. *Monthly Labor Review,* 1960, **83,** 925–932.

HARDIN, E. Characteristics of participants in an employee suggestion plan. *Personnel Psychology,* 1964, **17,** 289–303.

JACOBSON, E., TRUMBO, D., CHEEK, G., & NANGLE, J. Employee attitudes toward technological change in a medium sized insurance company. *Journal of Applied Psychology,* 1959, **43,** 349–354.

JACOBSON, E. H. The effect of changing industrial methods and automation on personnel. *Proceedings of the April 1957 symposium on preventive and social psychiatry.* Washington, D. C.: Walter Reed Army Institute of Research, 1959. Pp. 235–248.

LAWRENCE, P. R. How to deal with resistance to change. *Harvard Business Review,* 1954, **32**(3), 49–57.

LAWRENCE, P. R. *The changing of organizational behavior patterns: A case study of decentralization.* Boston: Harvard Business School, 1958.

LIPPITT, R., WATSON, J., & WESTLEY, B. *The dynamics of planned change: A comparative study of principles and techniques.* New York: Harcourt, Brace, 1958.

MCMURRAY, R. N. The problems of resistance to change in industry. *Journal of Applied Psychology,* 1947, **31,** 589–593.

MILLER, B. *Gaining acceptance for major methods changes.* New York: American Management Association, 1960.

NANGLE, J. E. A study of certain aspects of organizational communications within a medium-sized insurance company undergoing technological change. Unpublished doctoral dissertation, Michigan State University, 1961.

PAGÈS, M. The sociotherapy of the enterprise: The conditions of psychosocial change in industrial concerns and the role of the social psychologist as an agent of social change. *Human Relations,* 1959, **12,** 317–334.

RONKEN, H. O., & LAWRENCE, P. R. *Administering changes: A case study of human relations in a factory.* Boston: Harvard University Press, 1952.

SCHLEH, E. C. Selling technological changes as the company's way of life. *Personnel,* 1960, **37**(4), 57–66.

SELEKMAN, B. M. Resistance to shop changes. *Harvard Business Review,* 1945, **24**(Autumn), 119–132.

STEWART, M. Resistance to technological change in industry. *Human Organization,* 1957, **16**(3), 36–39.

STROUD, A., & ZELLNER, A. Program for computing efficient regression estimates and associated statistics. Systems Formulation and Methodology Workshop Paper 6204, Social Systems Research Institute, University of Wisconsin, May 22, 1962. (Mimeo)

TRUMBO, D. A. An analysis of attitudes toward change among the employees of an insurance company. Unpublished doctoral dissertation, Michigan State University, 1958.

TRUMBO, D. A. Individual and group correlates of attitudes toward work-related change. *Journal of Applied Psychology,* 1961, **45,** 338–344.

VANDER ZADEN, J. W. Resistance and social change. *Social Forces,* 1959, **37,** 312–315.

ZANDER, A. Resistance to change—its analysis and prevention. *Advanced Management,* 1950, **15**(1), 9–11.

ZELLNER, A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association,* 1962, **57,** 348–368.

ZELLNER, A. Estimators for seemingly unrelated regression equations: Some exact finite sample results. *Journal of the American Statistical Association,* 1963, **58,** 977–992.

(Received February 14, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 28–34

VOCATIONAL INTERESTS OF COMPUTER PROGRAMMERS

DALLIS K. PERRY AND WILLIAM M. CANNON [1]

System Development Corporation, Santa Monica

The revised SVIB was administered to 1,378 computer programmers. Primary analyses were limited to 1,003 males with at least 2 yr. of programming experience, whose jobs were primarily nonsupervisory, and who indicated satisfaction with programming. Programmers differ from other professional men primarily in their greater interest in problem solving, mathematics, and mechanical pursuits, and their lesser interest in people. Their interests are most similar to optometrists, chemists, engineers, production managers, mathematics-science teachers, and senior CPAs; but none of these existing keys adequately represents the interests of programmers. A Programmer key developed on half the sample and evaluated on the remaining half discriminates well between programmers and men in general. Satisfied programmers score significantly higher on the key than dissatisfied programmers.

Data processing is a rapidly growing occupation based on one of mankind's most significant technological developments—the electronic digital computer—and it promises to become one of the more important occupational fields. Because it is so new, data processing is not a well-known and widely understood field with the familiarity and popular stereotype that many other occupations have acquired. At best, such popular conceptions of occupations are frequently inadequate as grounds for vocational choice, but the data-processing field lacks even this much reality for many people. Consequently the existence of a measure that can be used to direct the attention of appropriately qualified persons to the data-processing field should be especially valuable for counseling, guidance, and recruiting. Exploratory studies by Sweetland (1962) and by Reinstedt, Hammidi, Peres, and Ricard (1964) have provided evidence of the feasibility of developing an interest scale for computer programmers. The present report describes the interests of computer programmers in terms of Strong Vocational Interest Blank (SVIB) scores and describes the development of an SVIB key for measuring interest in computer programming.

[1] The authors wish to acknowledge the assistance of D. P. Campbell in planning this study and in providing men-in-general response percentages for use in constructing the experimental interest keys; of J. M. Adams, H. R. Chandler, R. W. Rector, and Gloria Silvern in obtaining research participants; and of Mary Korman and R. L. McCornack in processing and analyzing the data.

METHOD

Subjects. Subjects (*Ss*) of the study were 1,378 computer programmers obtained from three sources. The largest number (1,170) came from 91 business, industrial, educational, and governmental organizations that participated in a 1963 salary survey of data-processing personnel. From 264 organizations represented in the survey, 130 were selected systematically to represent a cross section of organization types and data-processing staff sizes and were asked to participate in the study. A total of 91 organizations agreed and were sent research materials for 1,310 *Ss*. All contacts with individual *Ss* in these organizations were made through research coordinators in each organization. Responses were anonymous, but feedback of results to participants was made possible by a code-number list maintained by each coordinator. Of the 1,310 sets of materials distributed, 1,170, or 89%, were completed and returned.

A second group of 117 programmers was obtained through the Association for Computing Machinery (ACM). From the national ACM membership roster a sample of 208 members of 1- and 2-years standing was obtained. These members were invited to participate in the project, and the 137 who agreed to do so were sent the research materials directly by mail. The 117 complete returns represent a return rate of 85% of those who agreed to participate and 56% of those originally solicited.

A third group of 91 *Ss* was obtained in a similar way from the certification roster of the Data Processing Management Association (DPMA). Of the 169 persons contacted through DPMA, 111 agreed to participate and 91 completed the materials, representing a return rate of 82% of those who agreed to participate and 54% of those contacted. The ACM and DPMA groups were not anonymous.

Research instruments. Each participant in the study completed the interim revision of the SVIB (Strong, 1962) and a background information questionnaire including items on age, sex, educational and work history, job characteristics, programming qualifica-

tions, and a statement of satisfaction with programming as an occupation. This information made it possible to understand more clearly the nature of the participating sample, to define the final experimental sample, and to examine variations in interests among subgroups of programmers. All data were collected on a single mark-sense answer sheet. Instructions for each participant included an explanation of the purpose of the project and directions for completing both the SVIB and the background data form and for returning them directly to the project office in a return envelope provided.

Sample characteristics. The data-collection procedures limited participation in the project primarily, but not exclusively, to bona fide computer programmers. Background data were used, therefore, to eliminate from the sample: (*a*) those who indicated less than 2 years of programming experience (78 cases); and (*b*) those who indicated their primary job duty to be instructing programmers (9 cases), managing a data-processing installation (27 cases), or operating computer equipment (3 cases). The total sample included 186 females, and these were excluded from the primary sample because of the considerable evidence of differences in the vocational interests of men and women (Strong, 1943). Some evidence suggests that limiting interest-key criterion groups to individuals who express satisfaction with their occupation will result in more effective keys (Perry, 1960). Consequently a simple indication of satisfaction with the programming field was requested on the background data form, and participants who indicated that they "probably" (81 cases) or "definitely" (15 cases) would prefer another field were excluded from the primary sample. Restrictions on experience, primary duty, sex, and job satisfaction reduced the primary sample from 1,378 to 1,003. Subgroups of *Ss* excluded from the sample by a single restriction were retained for separate study.

The *Ss* in the primary sample ranged in age from 20 to 63, with a median of 31.8, mean of 32.8, and standard deviation of 6.5. The age range of 25–44 included 90% of the sample. More than 90% of the programmers had attended college, and two-thirds were college graduates. Among those reporting a college major, more than twice as many indicated a major in mathematics as in any other single field. Next most common were business administration and engineering. All but 10% of the programmers reported some kind of formal training in programming, the average amount being 174 hours, equivalent to 1 month of full-time training. About 90% had 8 years or less experience, reflecting the recency of development of the programming field. More than half were employed in industrial organizations, and more than two-thirds of those from known organizations were members of staffs of at least 50 programmers. Half of the programmers were engaged primarily in business programming, which is somewhat surprising in view of the preponderance of mathematical and scientific educational backgrounds. Another fourth of the programmers were doing primarily scientific programming, and the remainder were divided equally among combinations of business and scientific applications, military systems, and other applications, primarily nonmilitary software systems. More than half the programmers in the sample were receiving between \$700 and \$900 per month, and the median salary was exactly \$10,000 per year.

Procedure. The answer sheets were scored on 59 existing SVIB keys, and each participant was sent a profile of his individual scores. Similarity of programmers' interests to those of previously keyed occupations was assessed in terms of mean scores of the primary programmer sample on these keys.

Four experimental programmer interest keys were constructed from item responses differentiating a programmer criterion group (a randomly selected half of the primary programmer sample, $N = 500$) from a reference group representing an equal sampling of all of the criterion groups on which existing SVIB occupational keys are based. The remaining half of the primary programmer sample ($N = 503$) was used to evaluate the validity of the keys.

From the findings of Clark (1961), Nash (1965), and Strong, Berdie, Campbell, and Clark (1964) it appeared that a vocational interest key should be composed of 40–80 items selected to emphasize heterogeneity of content and scored with unit weights. In addition, Campbell[2] has suggested that better keys may result from scoring "Like" and "Dislike" responses to keyed items in opposite directions regardless of whether both responses significantly differentiate criterion and reference groups. The following keys were therefore developed as experimental programmer interest keys:

Key A. Composed of all responses to which the criterion- and reference-group percentages differed by 15% or more. Responses with larger criterion-group percentages are scored +1, and those with larger reference-group percentages are scored −1.

Key B. Composed of all items for which the criterion- and reference-group percentages of either "Like" or "Dislike" responses differed by 15% or more and for which the percentage difference for the "Indifferent" response was the smallest of the three. Responses with larger criterion-group percentages are scored +1, and those with larger reference-group percentages are scored −1, regardless of whether both responses to an item met the 15% difference criterion.

Key C. Composed of all responses to which the criterion- and reference-group percentages differed by 12% or more. Responses are scored as in Key A. Although separately analyzed, this key was developed primarily as an item pool for Key D.

Key D. Following the procedure outlined by Clark (1961), a sample of 500 men in general was scored on Key C; the upper and lower 27% of scores were identified; and the percentage differences in the responses of these two groups were computed for each item in the pool and used as indexes of response homogeneity. A bivariate plot of item statistics was prepared, with validity index (percentage difference between programmers and men in general) on one

[2] D. P. Campbell, personal communication.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF COMPUTER PROGRAMMERS' SCORES ON EXISTING SVIB KEYS

Key	*M*	*SD*
Optometrist	40.95	11.59
Physical Therapist	33.32	11.50
Dentist	26.80	10.71
Osteopath	27.04	09.69
Veterinarian	20.95	10.88
Physician	35.04	11.98
Psychiatrist	34.74	10.64
Psychologist	34.21	11.42
Biologist	35.64	11.54
Architect	31.42	11.09
Mathematician	31.55	12.06
Physicist	27.42	14.31
Engineer	41.81	11.29
Production Manager	39.92	08.66
Army Officer	35.78	12.58
Airplane Pilot	36.83	10.64
Carpenter	24.11	12.65
Forest Service Man	25.16	11.97
Farmer	32.31	09.90
Industrial Arts Teacher	29.34	10.80
Math–Science Teacher	40.45	09.22
Printer	29.93	09.41
Policeman	28.55	08.86
YMCA Physical Director	24.00	12.17
Personnel Director	35.00	11.94
Public Administrator	38.86	09.87
Rehabilitation Counselor	31.69	10.84
YMCA Secretary	20.90	12.58
Social Worker	28.96	12.39
Social Science Teacher	27.35	11.42
School Superintendent	27.66	12.14
Minister	21.60	11.96
Librarian	31.99	12.18
Artist	24.86	10.93
Art Teacher	22.14	10.85
Musician Performer	37.60	10.02
Music Teacher	30.06	11.00
C.P.A. Owner	36.59	09.62
Senior C.P.A.	42.45	08.93
Accountant	36.38	10.54
Office Worker	35.11	10.72
Credit Manager	35.38	11.95
Chamber of Commerce Executive	33.80	10.97
Business Education Teacher	31.70	11.82
Purchasing Agent	31.04	10.51
Banker	23.49	09.46
Pharmacist	28.90	09.09
Mortician	25.29	09.06
Sales Manager	25.75	10.07
Real Estate Salesman	29.54	07.37
Life Insurance Salesman	21.48	09.67
Advertising Man	28.42	08.51
Lawyer	32.32	09.37
Author–Journalist	28.64	08.62
President—Manufacturing Concern	25.39	10.95
Specialization Level	45.75	09.55
Occupational Level	58.44	06.85
Masculinity–Femininity	49.53	08.42

axis and homogeneity index (percentage difference between upper and lower groups of scores on Key C) on the other. A line drawn through the plot divided item responses with moderate validity indexes and high homogeneity indexes from those with very high validity indexes or with moderate validity indexes and low homogeneity indexes. The former responses were eliminated from the key, and the latter were retained. As examples of the effects of this process, all responses with validity indexes larger than 25 were retained; items with validity indexes of 12–18 were eliminated if their homogeneity indexes were about 30 or above but retained if they were about 20 or below; items with validity indexes of 20–25 were eliminated if their homogeneity indexes were about 50 but retained if they were about 30.

RESULTS

Interests of computer programmers. Means and standard deviations of the primary sample on 59 existing SVIB keys are given in Table 1, and the means are shown in profile form in Figure 1. The group has no very high mean scores—none in the A range. It is apparent that no one of these existing keys adequately represents the interests of computer programmers. Neither are there any extreme reject scores; only two of the eight mean scores in the C range fall below the middle third of men in general. Six of the mean scores—Optometrist, Chemist, Engineer, Production Manager, Math-Science Teacher, and Senior CPA —fall in the B+ range; and there are additional clusters of fairly high scores in the biological science and business detail groups, on Personnel and Public Administrator, on Pilot and Army Officer, and on the Musician key. The rather low scores on both Mathematician and Physicist contrast with the heavy emphasis on college training in mathematics among the programmers in the sample. These two keys were based on mathematicians and physicists listed in *American Men of Science,* including college professors. The much higher scores on Engineer and Chemist suggest that the mathematical and scientific interests of programmers are definitely applied rather than theoretical. The slightly elevated mean score on the Musician key provides some, but not very strong, support for the prevalent belief in a relationship between programming and musical activity. In general the profile of this group is similar to that obtained by Reinstedt et al. (1964), the primary difference being a larger number of

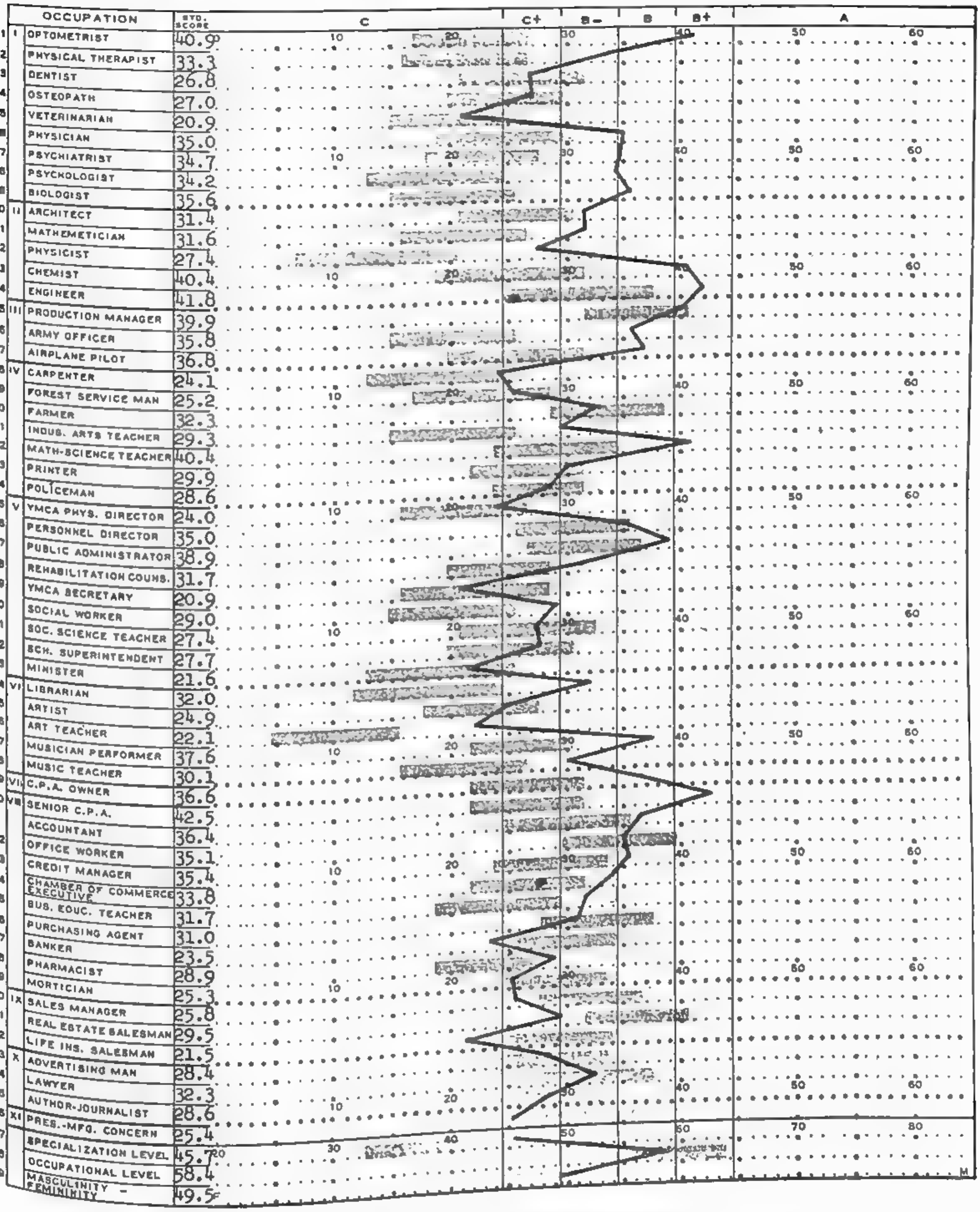


FIG. 1. SVIB profile for computer programmers.

fairly high scores—6 "B+" versus 1, 12 "B" versus 8—in the present sample. The higher scores occur primarily in Groups II (Physical Sciences) and III (Technical Supervision), with a few also in Group VIII (Business).

The interest items on which programmers differ most from professional men in general may be grouped into a few major categories that present a reasonably consistent picture and represent most of the differentiating items. It should be emphasized that the following description is a brief summary of ways in which programmers differ from other professional men and does not represent in any way the ab-

TABLE 2

CHARACTERISTICS OF EXPERIMENTAL PROGRAMMER INTEREST KEYS

Key	No. items	No. responses	*M*			*SD*		
			Criterion	Validation	Men in general	Criterion	Validation	Men in general[a]
A	66	102	23.93	24.03	3.17	8.41	8.08	10.50
B	51	102	18.65	18.70	−0.14	8.20	7.60	9.88
C	99	166	25.56	25.55	−3.44	11.12	11.31	14.71
D	72	106	18.74	18.74	−0.58	6.82	7.09	9.22

[a] Estimated.

solute amount of their interest in any area. Perhaps the most striking characteristic of programmers is their interest in problem- and puzzle-solving activities. Interestingly enough, this interest is not limited strictly to mental problems, as represented by a strong interest in all forms of mathematics, but extends as well to the mechanical area. Programmers show some liking for research activities, but no more interest than other men in most of the sciences themselves. Perhaps research work is attractive primarily as an opportunity to avoid routine, which appears to be particularly distasteful to programmers, as evidenced by their dislike for regimentation and their preference for varied and even risky activities. Another striking characteristic of programmers is their disinterest in people. Compared with other professional men, programmers dislike activities involving close personal interaction. They prefer to work with things rather than with people. They also indicate dislike of certain aesthetic activities, and it is noteworthy that they do not differ greatly from men in general in their responses to any of the "musical" items in the SVIB.

Experimental programmer keys. Statistical characteristics of the experimental programmer keys are shown in Tables 2 and 3. Keys A and D include about 100 scored responses, and the numbers of scored items are near the middle of the desired range. Key B contains fewer items, but inasmuch as both responses are scored for each item, the total number of responses is the same as for Key A. Key C, of course, contains more than the desirable number of items because it is merely a preliminary item pool to be reduced by further analysis. Note, however, that Keys A and C are nearly identical in performance despite the inclusion of 50% more items with lower validities in Key C. All four keys, in fact,

TABLE 3

CRITERION, VALIDATION, AND MEN-IN-GENERAL GROUP DIFFERENCES ON EXPERIMENTAL PROGRAMMER INTEREST KEYS

Key	Percentage of overlap with men in general[a]		Criterion group versus validation group				
			M difference		Variance difference		% overlap
	Criterion	Validation	*CR*	*p*	*F*	*p*	
A	27	26	0.19	>.10	1.09	>.10	99.5
B	30	28	0.10	>.10	1.16	<.10	99.8
C	26	27	0.04	>.10	1.04	>.10	99.9
D	23	24	0.00	>.10	1.08	>.10	100.0

[a] Using estimated men-in-general variance.

are similar in their performance, and all provide highly significant differentiation between programmers and men in general.[3] Criterion and validation groups overlap more than 99% on all keys, showing practically no shrinkage in validity. Key D is slightly the best of the four keys, but Keys A and C are nearly as good. Because considerable additional normative data on Key A, which was available earlier than Key D, were computed for inclusion in the SVIB manual, Key A is recommended as the Computer Programmer Interest Key. Scoring weights and normative data have been forwarded to the major SVIB scoring services. The letter-grade conversion table based on the validation group is given in Table 4.

TABLE 4

LETTER-GRADE CONVERSION TABLE AND NORMS FOR COMPUTER-PROGRAMMER INTEREST KEY

Letter grade	Raw score	Standard score	Percentage of distribution	
			Validation group	Revised men in general[a]
A	20–56	45–90	73.4	5.9
B+	16–19	40–44	11.1	6.2
B	12–15	35–39	8.4	9.4
B−	8–11	30–34	4.4	12.6
C+	4–7	25–29	2.0	14.7
C	−42–3	−32–24	0.8	51.2
	−16.3	0		

[a] Estimated.

Programmer scores of excluded subgroups. SVIB responses of the four major subgroups that were excluded from the primary programmer sample—females, inexperienced, and dissatisfied programmers, and those whose major responsibility was not a programming task—were scored on the Programmer key. Results are summarized and compared with the validation group in Table 5. Average scores of inexperienced programmers and of programmers whose major duties were not programming tasks do not differ significantly from the validation group, but average scores of both females and dissatisfied programmers are substantially below the validation group.

The lack of differentiation of the inexperienced programmers is not surprising, but somewhat lower scores for the group not engaged primarily in programming might have been expected. Undoubtedly many in this group were programmers and were continuing to do considerable programming, even though they regarded machine operation, training, or installation management as their major responsibility. It is possible, also, that the Programmer key is representative of the broad field of information processing as well as of the programming specialty within the field.

The lower scores of the dissatisfied programmers are considered additional strong evidence of the validity of the Programmer key. Although this group is more similar in its interests to satisfied programmers than to any of the other SVIB occupational groups, its dissatisfaction with the field is clearly re-

[3] Mean scores for men in general were obtained by summing the cross-products of men-in-general response percentages and corresponding item weights, rather than by scoring the men-in-general group. Men-in-general standard deviations were estimated as 1.3 times the validation-group standard deviations, on the basis of the standard-score standard deviations reported for men in general on existing SVIB keys (Strong, 1959).

TABLE 5

SCORES OF EXCLUDED SUBGROUPS ON THE COMPUTER-PROGRAMMER KEY

Group	*N*	*M*	*SD*	Difference from validation group				
				M diff.	*t*	*p*	Variance ratio	*p*
Females	169	20.17	7.93	3.86	5.40	<.001	1.02	>.10
Inexperienced	49	25.73	8.48	−1.70	−1.40	>.10	1.05	>.10
Nonprogram.	33	22.73	8.79	1.30	0.89	>.30	1.09	>.10
Dissatisfied	76	19.88	9.59	4.15	4.07	<.001	1.19	>.10

flected in its low average Programmer score. Because the members of this group have in common their employment as programmers and the interests that drew them into the programming field, whereas their dissatisfactions with the field and their other interests are likely to be widely varied, it is not surprising that no other single key represents their interests as well as the Programming key.

The interests that differentiate men from women in general apparently differentiate female programmers from male programmers as well; and the difference is sufficient to require separate norms if the men's key is to be used for women. It appears desirable as well to construct a computer-programmer key for the SVIB for Women, using women in general as the reference group.

REFERENCES

CLARK, K. E. *The vocational interests of nonprofessional men.* Minneapolis: University of Minnesota Press, 1961.

NASH, A. N. A study of item weights and scale lengths for the SVIB. *Journal of Applied Psychology,* 1965, **49**, 264–269.

PERRY, D. K. Problems of item form and criterion group definition. In W. L. Layton (Ed.), *The Strong Vocational Interest Blank: Research and uses.* Minnesota Studies in Student Personnel Work No. 10. Minneapolis: University of Minnesota Press, 1960.

REINSTEDT, R. N., HAMMIDI, B. C., PERES, S. H., & RICARD, E. L. Computer Personnel Research Group programmer performance prediction study. *Memorandum RM-4033-PR.* Santa Monica: RAND Corporation, 1964.

STRONG, E. K., JR. *Vocational interests of men and women.* Stanford: Stanford University Press, 1943.

STRONG, E. K., JR. *Strong Vocational Interest Blanks for Men and Women.* Palo Alto: Consulting Psychologists Press, 1959.

STRONG, E. K., JR. Good and poor interest items. *Journal of Applied Psychology,* 1962, **46**, 269–275.

STRONG, E. K., JR., BERDIE, R. F., CAMPBELL, D. P., & CLARK, K. E. Proposed scoring changes for the Strong Vocational Interest Blank. *Journal of Applied Psychology,* 1964, **48**, 75–80.

SWEETLAND, A. Factors in selecting and training programmers. *Memorandum RM-3245-PR.* Santa Monica: RAND Corporation, 1962.

(Received February 21, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 35-38

COMPATIBILITY EFFECTS IN A TWO-HAND CRANKING TASK[1]

DON TRUMBO, STEVEN ROGERS, AND LLOYD L. AVANT

Kansas State University

48 Ss were run on a 2-hand cranking task in 4 independent groups with different control-display linkage. Within-group conditions included 3 stimulus codes and 4 response combinations. Total adjustment time and response latencies for each hand served as response measures. Results for total adjustment time showed significant Stimulus Code and Practice effects, but no differences for Response Combinations or for the Stimulus Code-Response Combination interaction. Latency data revealed a pronounced Stimulus Code effect consistent with that for total adjustment time and demonstrated that control of the lateral dimension yielded shorter latencies than the in-out dimension, regardless of the hand involved. Thus, there was evidence of S-R but not R-R compatibility effects.

Cross, Noble, and Trumbo (1964) have reported an attempt to assess compatibility effects, including both stimulus-response (S-R) and response-response (R-R) effects. In that study, which involved a two-hand cranking task, it was found that performance was a function of the stimulus code employed (S-R compatibility), but was relatively independent of the combination of responses required (R-R compatibility), even though the latter included both symmetrical and identical movements of the two cranks. Thus, while demonstrating S-R effects, the results were in contrast to the classic principles of time and motion (Barnes, 1958) and with the concept of R-R compatibility, as discussed by Fitts and his associates (Fitts, Noble, Bahrick, & Briggs, 1959), both of which would seem to predict an optimal combination of movements of the two hands.

Two limitations of the prior study prompted the present research. First, only six subjects (*Ss*) were employed in a repeated measures design in the previous study. Second, and more important, the control-to-pointer linkage was fixed and consistent with population stereotypes. That is, for all conditions a clockwise turn of the right-hand crank moved the pointer (controlled element) to the right, and a clockwise turn of the left-hand crank moved the pointer away from *S*. Therefore, even though successive trials required all four combinations of movements of the two hands, all such combinations were consistent with population stereotypes, a factor which may have masked compatibility effects based on symmetry or identity of responses.

METHOD

Subjects. There were 48 university males who served as *Ss* and received research participation credit in introductory psychology courses for their services.

Apparatus. The apparatus described in detail in the prior report (Cross et al., 1964) was used, with minor modifications, in the present study. Basically, the system consisted of a 15-inch × 15-inch rear projection screen mounted just beyond a 10 × 10 horizontal matrix of target points and opposite *S*'s control cranks. The target board consisted of $\frac{3}{8}$-inch-diameter metallic contact points, $1\frac{3}{4}$ inch apart, on a dull black nonconducting background. Matrix columns and rows were coded from A to J and 1 to 10, respectively. The controls were identical lightweight cranks in the vertical plane with a center-to-center distance of 15 inches and radii of $6\frac{1}{4}$ inch. One revolution of the crank resulted in 1 inch of pointer movement. Control linkage was mechanical and consisted of gears, sprockets, and bicycle chains. The pointer was located at the vertex of a plexiglass wedge, the base of which was attached to the chain linkage. Contact of the pointer with the correct target stopped a "Total Adjustment Time" clock, which started with the activation of a slide projector and, consequently, with the onset of the stimulus.

Continuous response records were approximated by having a brush, located on the back of each crank, sweep 16 contact points which were wired in pairs to 16 pens on an Esterline-Angus Recorder. The wiring code was such that a half revolution of one crank would result in a cascade of pips across eight of the pens, the slope of the cascade reflecting the direction

[1] This research was supported, in part, by the Kansas State University Faculty Research Fund.

and rate of crank movement. Contact points at 10:00 o'clock and 2:00 o'clock positions were wired to a clock for each crank. These measured the latencies for 120-degree crank movement, since all trials were initiated from a 6:00 o'clock crank position.

Stimuli. The stimulus codes, identical with those used in the prior study, were: (*a*) *Instructional* (I), wherein, for example, "R-5, U-3" instructed *Ss* to move to the target "right five and up three" from his resting position; (*b*) *Digital* (D), which presented the alpha-numeric coordinates of the target position, for example, "J-5," and (*c*) *Pictorial* (P), which consisted of a diagrammatic representation of the 10 × 10 matrix with the correct target designated by a large black square. For each stimulus type, 36 slides were prepared. The sequences of slides were ordered so that (*a*) each of the four clockwise-counterclockwise response combinations of the two hands (CC-C, CC-CC, C-C, C-CC) would occur with equal frequency throughout the series, (*b*) each response combination would occur once, randomly positioned, within each block of four trials, and (*c*) the total distance to be traveled would be equal for all response conditions. The series of 36 target positions were identical for all three stimulus codes.

Design and procedures. The 12 *Ss* assigned to any one of the four control-to-pointer (C/P) conditions were subsequently assigned, two each, to one of the six orders of the three stimulus codes, wherein they received all 36 stimuli from each code before proceeding to the next code. Thus, stimulus codes and response combinations were balanced within *Ss*, while C/P conditions were varied for independent groups of *Ss*.

The four C/P groups were run successively, rather than concurrently, because of the complexity of the alterations necessary to change the C/P conditions. However, force requirements to move the cranks were kept constant throughout all conditions.

Individual *Ss* were instructed in the task and in the stimulus codes and received three practice trials on each code before test trials began. They were told that the experimenter (*E*) would give a verbal "ready" signal about 3 seconds before the onset of each stimulus and that a red signal light would come on and the stimulus would go off when they made contact with the correct target. Speed was emphasized, but no response set (such as: "be sure to turn both cranks simultaneously") was given. All 108 trials were run in one session with approximately 15 seconds between successive trials while *E* recorded clock times and reset his control panel to complete the circuit for the target position of the next trial.

Results

Total adjustment time. Table 1 presents a summary of the analysis of the total-adjustment-time data. Mean values for the four groups (C/P conditions) were nearly identi-

TABLE 1

Summary of Analyses of Variance for Total Adjustment Time and for Left- and Right-Hand Latencies

Source	Total adjustment time			Left hand[a]		Right hand[a]	
	df	*MS*	*F*	*MS*	*F*	*MS*	*F*
Control-to-pointer groups (G)	3	15.74	.22	37.68	8.74**	36.15	15.86**
Instructional-Digital-Pictorial Sequences (S)	5	39.33	.56	2.90	.67	.29	.13
G × S	15	79.60	1.14	6.28	1.46	7.79	3.43**
Ss/G × S	24	69.70	—	4.31	—	2.28	—
Stimulus Code (St)	2	2786.62	139.61**	107.79	99.80**	120.35	138.65**
Runs of 36 (R)	2	440.14	22.05**	4.05	3.75*	1.06	1.22
G × St	6	39.78	1.99	10.42	9.65**	11.37	13.10**
G × R	6	7.52	.37	.65	.60	1.55	1.79
Residuals (*Ss*/Seq)	80	19.96	—	1.08	—	.87	—
Blocks (B)	8	131.64	1.75				
Response Combinations (RC)	3	28.13	.37	1.76	.14	2.00	.18
B × R	16	16.33	.21				
RC × S	6	14.69	.19	.62	.05	1.89	.17
B × RC	24	34.64	.46				
Pooled Residuals	4,983	75.33	—	12.25 (423 *df*)	—	11.21 (423 *df*)	—

[a] Analyses for left and right hand used RC means across the 9 four-trial blocks, hence all terms containing "Blocks" disappear and the *df* for pooled residual is reduced.

* $p < .05$.

** $p < .01$.

cal with F $(3/24) = .225$, $p > .05$. However, Stimulus Codes (St) and Runs (R) of 36 stimuli were significant sources of variance. The St effect was the result of poor performance on the I Code ($\bar{X}_I = 9.79$ seconds), since the P and D codes did not differ ($\bar{X}_P = 7.56$; $\bar{X}_D = 7.63$). The Runs effect indicates a general practice effect with a significant decrease in total adjustment time between the first and second series of 36 stimuli ($\bar{X}_{1,2} = 8.89$ and 8.17 seconds, respectively), but not between the second and third series ($\bar{X}_3 = 7.92$). F tests of the Response Combinations (RC) and the RC $\times$ St interaction effects were nonsignificant. Among the remaining sources of variance only the Groups $\times$ Stimuli (G $\times$ St) interaction approached significance with F $(6/80) = 1.99$, $.10 < p > .05$. Collectively these results indicated that Stimulus Code, but not Response Combinations, or Response Combinations in interaction with Stimulus Code, contributed significantly to variance in total adjustment time.

Response latencies. Summaries of the analyses of variance for the response latencies of left and right hands are also shown in Table 1. The results for the two hands are similar in that the Groups, Stimuli, and Groups $\times$ Stimuli sources of variance were significant, but differ in that the Runs effect was significant for the left hand, while the Groups $\times$ Sequences interaction was significant for the right hand.

These effects are illustrated in Figure 1 for both hands. It is apparent that the Group effect is reciprocal for the two hands and a function of the dimension controlled by each hand. That is, in Groups 1 and 2 the right hand controls the lateral movement of the pointer, whereas in Groups 3 and 4 the left hand controls the lateral movement, and, in each case, control of the lateral dimension is associated with shorter latencies. The Stimulus effect, as shown in Figure 1, is the result of particularly long latencies on the Instructional stimuli for both hands. This is consistent with the results from the total-adjustment-time analysis, as well as those reported by Cross et al. (1964). Finally, Figure 1 shows that the Groups $\times$ Stimuli interaction results primarily from the fact that latencies

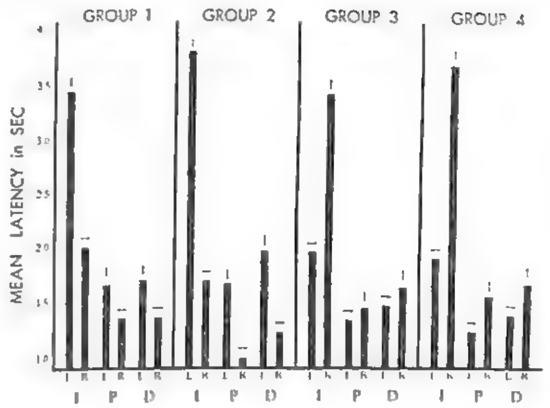


FIG. 1. Mean latencies, by control-to-pointer group, for each hand (L = Left, R = Right), with each Stimulus Code (I, P, D). (Arrows indicate the direction controlled by each hand.)

for control of the inward-outward, or Y, dimension (Groups 1 and 2, left hand; Groups 3 and 4, right hand) are relatively greater with the Instructional than with the other two codes, and that under the "I" code there is a greater difference between the latencies of the two hands for any given group.

DISCUSSION

The results with the Total-Adjustment-Time criterion fail to demonstrate either R-R or S-R-R compatibility effects. Simple R-R effects would have been manifested in a significant Response Combination effect with, for example, either symmetrical (C-CC or CC-C) or identical (C-C or CC-CC) response combinations yielding shorter adjustment times. Similarly, S-R-R compatibility effects would have resulted in a significant interaction between Stimulus Codes and Response Combinations, indicating that the optimum combination of responses was a function of the manner of encoding the input information. The lack of such effects indicates that, with the present task and with the criterion of total adjustment time, neither the symmetry-identity nor the S-R-R ensemble conditions affect task performance.

Furthermore, failure to find significant differences among the four groups with different control-display linkages does not support the expectation that outcome performance would reflect differences in population stereotypes with respect to C to D relations. Thus, only

Stimulus Codes and Runs (or practice) were found to affect the Total Adjustment Time "outcome" criterion. The former effect indicates that the Instructional code, which conveys the same amount of information as the other two codes, results in a less compatible S-R ensemble than either the Pictorial or Digital codes. The Runs effect simply indicates a positive practice effect over the three runs of 36 stimuli.

Only one source of evidence suggests a type of S-R-R compatibility; the Groups × Stimulus interaction, which approached significance. This effect, were it more reliable, would indicate that efficiency of a particular stimulus code depends on the crank-to-pointer linkage. On the other hand, if one conceives of the stimuli and the display as two stimulus elements, one might consider such an effect as an S-S-R form of compatibility.

The latency data for the two hands provide some evidence to account for the large Stimulus Code effect in the total-adjustment-time analysis. Combined latencies for the two hands are much greater on the Instructional (I) code than on either of the remaining codes. Furthermore, differential latencies for the two hands are greatest for the I code and show the greatest effect of direction of control (lateral versus in-out movement of the pointer). These results suggest that when *Ss* are faced with the I code, they follow the instructions explicitly and sequentially, namely, "R-5, U-3" is responded to by first moving the pointer five units to the right, then moving it up three units, rather than moving both controls simultaneously. Such a response pattern results in a greater travel distance and longer adjustment times. This conclusion was supported by visual inspection of the continuous records which revealed little temporal overlap between responses of the two hands. It may also be true that the decoding of the I stimuli is more complex, since "R" and "U" must be translated into "right" and "up" which, in turn, must be related to directions on the display matrix. Furthermore, the instructions provided by the I code are meaningless unless *S* maintains his starting position as a point of reference. This may be more easily accomplished by the strategy of following one dimension at a time. It was observed that *S* occasionally lost his reference with the I code, in which case the trial was terminated unless the target was attained in 15 seconds.

Finally, information was ordered spatially on the I slides in that the left-right code was presented as one line with the up-down code immediately below as a second line. Conceivably, this would induce the observed response pattern, although a similar fixed order occurred with the Digital code in that the alpha code always preceded the numeric (J-5, never 5-J), without comparable effects.

REFERENCES

BARNES, R. M. *Motion and time study.* New York: Wiley, 1958.

CROSS, K., NOBLE, M., & TRUMBO, D. On response-response compatibility. *Human Factors,* 1964, 6, 31–37.

FITTS, P. M., NOBLE, M., BAHRICK, H. P., & BRIGGS, G. E. Skilled performance: Parts I and II. USAF WADC Final report, 1959, No. AF 41 (657)-70.

(Received February 21, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 39–45

WAGE INEQUITY AND JOB PERFORMANCE: AN EXPERIMENTAL STUDY

I. R. ANDREWS

University of California, Berkeley

96 students worked at 1 of 3 piece rates: equitable wage (20¢), underpay (15¢), or overpay (30¢). Half of these students worked on an inherently interesting task while the other half worked on an inherently dull task. As predicted, underpaid Ss maintained equity by increasing work quantity at the expense of work quality, whereas overpaid Ss maintained equity by reducing work quantity and increasing work quality. Over and above the job-performance difference between piece-rate groups, there were pronounced differences within piece-rate groups as a function of individual differences in previous wage experience. 2 other hypotheses, 1 about task difference and 1 about the relative effects of underpay and overpay inequity, were not consistently supported by the data.

This study provides a further test of Adams' (1963a) theory of inequity, as applied to the special case of wage inequity. Inequity is said to exist when "inputs" (such as age, education, effort expended, etc.) are not in the expected balance with "outcomes" (such as pay received, job status, etc.). A perceived inequity is said to motivate the worker to correct the imbalance through the adjustment of inputs or outcomes, either in actual fact or through psychological distortion. (For a more precise statement of the theory, see Adams, 1963a, 1965.)

Experimental data (Adams, 1963b; Adams & Jacobsen, 1964; Adams & Rosenbaum, 1962; Arrowood, 1961) tend to support the theory, but are limited in the following ways: only college students have been studied; work periods have been very short; workers were isolated from all co-workers; only two work tasks have been used, and there has been no comparison across these two tasks; wage-inequity dissonance has been induced only through a manipulation of perceived inputs, specifically, the degree to which the worker was said to be qualified for the job; there may have been some confounding by the possibility that overpaid workers were striving harder to protect their jobs; only the overpayment side of inequity has been studied. Though also limited by the first three points above—college-student sample, short work period, and worker isolation—the present study attempted to go beyond earlier studies in the following respects:

1. Inequity dissonance was induced through a manipulation of outcomes instead of through a manipulation of perceived inputs.
2. Since all workers shared the expectation that their job would last for only 2 hours, there was no confounding by the possibility that some workers worked harder because they were trying to protect their job.
3. The effects of underpay inequity were examined. More specifically, the experiment attempted to compare a small amount of underpay inequity against a larger amount of overpay inequity.
4. The possibility of a task difference in the effects of wage inequity was tested.
5. Also studied was the effect of previous wage experience on a worker's reaction to his present wage.

In earlier studies, perceived job inputs (qualifications for the job) were varied by the experimenter (E) while wage rate was held constant. In the present study, wage rate was varied by E while perceived job inputs were held constant by random assignment of workers to different wage groups. This modification served three purposes: first, it tested the theory through a different kind of experimental manipulation; second, it eliminated the possibility of confounding due to differences in feelings of job security; and third,

it permitted a more precise manipulation of underpay and overpay inequity.

As in Adams' later studies, this study employed a piecework pay system. This choice was dictated in part by the fact that underpayment on an hourly pay basis would have led to a self-selection of subjects (*S*s), thus producing a biased sample.

Under the piecework pay system, if a worker perceives himself as underpaid, he can improve his outcomes through increased quantity of production, even though added speed may lead to a loss in work quality. Conversely, if he perceives himself as overpaid, he can reduce his outcomes by reducing production quantity, which gives him time to improve work quality. Therefore, under a piecework pay system, Hypothesis 1a can be stated as follows: As compared to workers assigned to an equitable piece rate, workers assigned to a lower piece rate will tend to produce more pieces of work but of lower quality, whereas workers assigned to a higher piece rate will tend to produce fewer pieces of work but of higher quality.

Stated in the above form, the hypothesis does not allow for individual differences in how a given pay rate is *perceived*. Yet, it seems reasonable to suspect that a worker's perception of his present pay rate will be influenced by the pay rates he has experienced on previous jobs. Therefore, as a corollary to Hypothesis 1a, Hypothesis 1b predicted differences *within* experimental groups as a function of previous wage experience: Within experimental groups, the more that present wage potential exceeds previous high wage, the lower the work quantity and the higher the work quality; the more that present wage potential is less than previous high wage, the higher the work quantity and the lower the work quality.

As defined by Adams, outcomes include not only pay received, but also such factors as intrinsic job interest, job status, etc. Moreover, outcomes are assumed to be additive and interchangeable, so that a partial absence of one outcome can be balanced by the presence of another outcome. For example, a partial absence of pay can be balanced by high intrinsic job interest. From this line of thought can be derived Hypothesis 2: Wage inequity will have more of an effect upon job performance and job attitudes when the task is dull than when the task is inherently interesting.

Adams states that it is probable that the thresholds for inequity are different in cases of under- and overcompensation. Like Jacques (1961), Adams believes that workers are more sensitive to undercompensation than to overcompensation. This idea was reflected in Hypothesis 3: A small underpay inequity will affect job performance as much as or more than a larger overpay inequity.

METHOD

There were 96 university students, hired through the student placement office, who were assigned randomly to one of six conditions of two tasks:

(*a*) Interviewing: Low pay rate, 15¢ per piece ($n = 16$); Equitable rate, 20¢ per piece ($n = 16$); High pay rate, 30¢ per piece ($n = 16$).

(*b*) Data Checking: Low pay rate, 15¢ per piece ($n = 16$); Equitable rate, 20¢ per piece ($n = 16$); High pay rate, 30¢ per piece ($n = 16$).

A comparison of these six groups on such factors as age, year in school, and previous wages earned showed no significant differences.

Procedure

Through pretesting, both tasks were adjusted to an average production of nine pieces per hour. Then, by dividing nine pieces of work into the average hourly pay rates recommended by pretest *S*s ($1.78 for interviewing and $1.79 for data checking), the equitable pay rate was set at 20¢ per piece. The overpay rate was set arbitrarily at 30¢ per piece and the underpay rate was set arbitrarily at 15¢ per piece. This asymmetric change from the 20¢ equitable rate was in keeping with the third hypothesis which stated that a small underpay inequity would have an effect equal to, or greater than, a larger overpay inequity.

That these differences in wage potential were perceived as intended was confirmed by differences in the size of the discrepancy between "worker recommendations for an appropriate hourly pay rate" and "the amount of money which had already been earned." On the average, the recommended hourly pay rate exceeded pay already earned by $0.13 an hour in the case of the 15¢ piece-rate group, as compared to $0.05 an hour in the case of the 20¢ piece-rate group. For the 30¢ piece-rate group, the recommended hourly pay rate was $0.69 an hour *less* than the pay already earned.

Over and above the between-group differences predicted by Hypothesis 1a, Hypothesis 1b predicted within-group differences as a function of previous wage experience. For the total sample, the

25 highest paid persons earned $2.25 or more per hour while the 25 lowest paid persons earned $1.55 or less per hour, leaving a middle group which ranged from $1.60 to $2.20 per hour. These three previous wage categories were used to divide workers within each of the piece-rate groups as required by Hypothesis 1b.

As required by the second hypothesis, one task had to be of greater inherent interest than the other task. For this purpose, interviewing students on campus was selected as the interesting task while checking transcribed numbers was selected as the dull task. As evaluated by both pretest and experimental *Ss*, the interviewing job received an average rating of about 6 on a 7-point scale, while the data-checking job received an average rating of less than 3. This difference was highly significant.

Aside from this difference in the degree of inherent job interest, the two tasks were made as similar as possible: each employee worked for only 2 hours and had no expectation for employment beyond the two hours; each employee was given 30 pieces of work with the instruction ". . . to work on these for only 2 hours"; by pretesting and adjusting as required, the two tasks were equilibrated for the average number of pieces completed within a specified time; in pretesting and in the experimental data, the performance variance was about equal for the two tasks; piece rates paid for the jobs were identical; specific instructions about work quantity and quality were omitted deliberately; only male employees were hired; each employee worked in isolation from other employees; both tasks were presented in the guise of legitimate jobs.

For each of the two tasks there were two dependent variables, work quantity and work quality. Work-quantity data were entered in raw score form, that is, the actual number of work pieces completed in 2 hours. This was possible because the two tasks were comparable in average rate of production, production variance, and in the shape of the frequency distribution (symmetric and somewhat platykurtic).

Work-quality data, on the other hand, were completely different for the two tasks and, therefore, had to be converted to standard score form for statistical comparisons. Interviewer work quality was defined as the average number of words recorded per interview—a measure which correlated over .80 with the average number of ideas recorded per interview. The higher the average number of words recorded, the higher the work quality. For the data checkers, work quality was defined as the number of transcription errors *missed* divided by the total number of errors which had been planted in those pages which the worker said he had checked. The smaller the proportion of errors missed, the higher the work quality.

RESULTS

As predicted by Hypothesis 1a and Hypothesis 1b, quantity of output was negatively related to assigned piece rate and positively related to previous high wage per hour (Table 1). The mean difference between the 15¢ and 30¢ piece-rate groups was significant at the .02 level ($t = 2.56$), while the mean difference between the $2.25-or-more and the $1.55-or-less groups was significant at the .01 level ($t = 2.91$).

Also as predicted, quality of work was positively related to assigned piece rate and negatively related to previous high wage per hour (Table 2). The mean difference between the 15¢ and 30¢ piece-rate groups produced a t value of 1.83, not significant with a two-tailed test. If one accepts a one-tailed test on the grounds that direction of effect was predicted, then this t value was significant at the .05 level. The mean difference between the $2.25-or-more and the $1.55-or-less groups was significant at the .05 level ($t = 2.33$).

When work quality and quantity are considered simultaneously, persons who feel overpaid should produce fewer pieces of higher quality work—while those who feel underpaid should produce more pieces of lower quality work. The proportion of workers who

TABLE 1

QUANTITY OF WORK AS A FUNCTION OF ASSIGNED PIECE RATE AND PREVIOUS WAGE EXPERIENCE

Previous high wage per hour	15¢ per piece ($n = 32$)	20¢ per piece ($n = 32$)	30¢ per piece ($n = 32$)	Row average
$2.25 or more ($n = 25$)	23.7	18.2	19.6	21.0
$1.60 to $2.20 ($n = 46$)	22.2	20.3	17.0	20.0
$1.55 or less ($n = 25$)	16.7	15.9	15.1	16.0
Column average	21.1	18.9	17.7	19.2

TABLE 2

QUALITY OF WORK AS A FUNCTION OF ASSIGNED PIECE RATE AND PREVIOUS WAGE EXPERIENCE

Previous high wage per hour	15¢ per piece ($n = 32$)	20¢ per piece ($n = 32$)	30¢ per piece ($n = 32$)	Row average
$2.25 or more ($n = 25$)	−0.29	−0.71	+0.18	−0.19
$1.60 to $2.20 ($n = 46$)	−0.15	−0.32	+0.20	−0.12
$1.55 or less ($n = 25$)	−0.01	+0.57	+0.62	+0.38
Column average	−0.16	−0.16	+0.30	0.00

fell in the former category—work quantity below the total sample median and work quality above the total sample median—is shown in Table 3. The proportion of workers who fell in the latter category—work quantity above the total sample median and work quality below the total sample median—is shown in Table 4. In both tables the influence of piece rate and the influence of previous high wage are readily apparent. For example, if a worker's previous high wage was \$1.55 or less and his present piece rate was 30¢, then there was a 75% chance that he would turn in fewer pieces of higher quality work and a 13% chance that he would do the reverse. The probabilities were strikingly different for a person whose previous high wage was \$2.25 or more and whose present piece rate was 15¢; for such a worker there was a 10% chance that he would turn in fewer pieces of high quality work and a 60% chance that he would turn in many pieces of lower quality work.

According to the second hypothesis, the effects of wage inequity should have been greater on the dull task (data checking) than on the interesting task (interviewing). As measured by mean differences in work quantity and work quality, the data did not support the hypothesis. On the other hand, two pieces of evidence did provide some encouragement for the task-difference hypothesis. First, the correlation between "work quantity" and "the size of the discrepancy between present wage potential (9.5 times the assigned piece rate) and previous high wage per hour" was −.54 for the data checkers but only −.32 for the interviewers. The comparable correlations for work quality were .35 for the data checkers but only .20 for the interviewers. Though not reaching statistical significance, these differences in correlation size were both in the direction predicted. A second kind of support for the hypothesis stems from the way job-attitude scores (on a postwork questionnaire) varied as a function of assigned piece rate. Among the data checkers there was a systematic change across piece-rate groups—the higher the piece rate, the more favorable the job attitude on five out of six scales. In contrast, job attitudes among interviewers remained fairly constant from one piece-rate group to another. Among interviewers, for example, the mean difference between the 15¢ and 30¢ piece-rate groups on the underpaid/overpaid scale was *not* significant at the .05 level; among data checkers, on the other hand, this difference was significant at the .001 level.

TABLE 3

CHANCE IN 100 THAT WORK QUANTITY WAS KEPT LOW WHILE WORK QUALITY WAS INCREASED

Previous high wage per hour	Assigned piece rate	Chance in 100
\$2.25 or more (n = 25)	15¢	10
	20¢	0
	30¢	40
\$1.60–\$2.20 (n = 46)	15¢	8
	20¢	21
	30¢	36
\$1.55 or less (n = 25)	15¢	44
	20¢	75
	30¢	75

TABLE 4

CHANCE IN 100 THAT WORK QUANTITY WAS INCREASED WHILE WORK QUALITY WAS SACRIFICED

Previous high wage per hour	Assigned piece rate	Chance in 100
\$2.25 or more (n = 25)	15¢	60
	20¢	40
	30¢	20
\$1.60–\$2.20 (n = 46)	15¢	46
	20¢	58
	30¢	29
\$1.55 or less (n = 25)	15¢	0
	20¢	13
	30¢	13

According to the third hypothesis, job-performance differences between the 15¢ and 20¢ groups should be equal to or greater than the performance differences between the 20¢ and 30¢ groups. As shown by the column averages in Table 1, work-quantity data were consistent with this hypothesis; the mean difference for the 15¢/20¢ com-

parison was slightly larger than the mean difference for the 20¢/30¢ comparison, 2.2 as compared to 1.2 pieces of work. Work-quality data, on the other hand, were not consistent with the third hypothesis; the mean difference for the 15¢/20¢ comparison was 0.00 as compared to a mean difference of 0.46 for the 20¢/30¢ comparison. Though not reaching statistical significance until the .10 level, the direction of this difference in mean differences was inconsistent with the hypothesis.

Discussion

The above results are consistent with earlier studies which showed that dissonance from overpayment (on a piecework pay system) can lead to a reduction in work quantity and an increase in work quality. The above results add to previous studies in the following ways:

1. Unlike earlier studies which produced inequity by the manipulation of perceived job inputs (qualifications for the job) the present study produced inequity by the manipulation of assigned piece rate. It was assumed that workers would check periodically to see how much money they were earning; those who perceived their earning rate as low would be tempted to increase their rate of production even though work quality had to be sacrificed and vice versa for those who found they were earning too much. Some support for this interpretation follows from the fact that 30¢-per-piece workers tended to do fewer pieces of work in their second hour, while 15¢-per-piece workers tended to do more pieces of work in their second hour.

Postwork discussions with workers revealed that another fairly common approach to the job was the establishment of a production quota before work was begun. For example, if a worker had a high previous wage of $2.00 per hour, then assignment to a 20¢-per-piece rate would lead to a work quota of 10 pieces per hour. That same worker assigned to a 30¢-per-piece rate would set a quota of about 7 pieces per hour, once again hoping to earn at a rate consistent with his previous wage experience. To the extent that such a worker could maintain his self-imposed rate of production, dissonance could have been avoided. For these workers, it could be argued that Adams' dissonance model was less appropriate than a cognitive consistency model which does not postulate dissonance, for example, Sampson's (1963) expectancy congruence model. This change in explanation, however, does not negate Adams' basic point, namely, that workers adapt their productive effort to maximize equity rather than total earnings.

In both of the above reaction patterns there was an adjustment in rate of output which reduced or prevented wage inequity. But not all persons responded to a perceived inequity by adjusting rate of output. For example, one 15¢-per-page data checker said that he just could not believe that it was real—that he must have misheard the instructions. He also said that at times he felt like going through the pages and changing many of the numbers. In spite of these misgivings, he turned in good quality work. Several of the underpaid workers rationalized their low earning rate by saying that they just were not well suited to the job. On the opposite side of the equity ledger, some of the overpaid workers rationalized their high earnings by saying that they were unusually skillful at this kind of work. Of these workers, several cited previous job experience which, as they saw it, developed the required skills.

Not included in any of the above reaction patterns (or in the results data) were three 15¢-per-piece workers who quit the job after 30 or 40 minutes, stating that this just was not their kind of work. No workers left their jobs in the 20¢- and 30¢-per-piece groups.

Also not included in the above reaction patterns were three 30¢-per-piece data checkers who tried to reduce overpay inequity by reporting fewer pages completed than they had actually done. Lastly, there were some workers who claimed that they were unaware of how much money they were earning. They said that they just did their jobs and did not watch the clock or count the pieces of work completed.

2. Whereas earlier studies considered only the overpayment side of inequity, the present

study also examined the effects of underpayment dissonance. As expected, the average work quantity was greater for the 15¢-per-piece workers (21.1) than for the 20¢-per-piece workers (18.9). However, the expected trend did not hold for work quality where there was no difference between the 15¢- and 20¢-per-piece groups. Several explanations for this lack of difference are possible: first, the 5¢-per-piece differential between the 15¢ and 20¢ groups was too small; second, increased speed on the monotonous data-checking task kept workers more alert with the result that work quality did not suffer from a slight underpayment; third, possible confounding by a change in personnel in the student placement office—an interpretation fostered by the fact that the expected trend was very much intact *before* the addition of the last 16 *S*s, workers who were hired after the personnel change at the placement office. In brief, *E* believes that the inconsistency of work-quality data should not be taken as evidence against the hypothesis. A subsequent study, using 10¢ per piece as the unfairly low wage and 25¢ per piece as the equitable rate, found that the underpaid *S*s not only produced more pieces of work (interviews), but did so by a statistically significant reduction in work quality (Lawler & O'Gara, 1966).

3. Perhaps the most interesting result in the present study was the effect of previous wage experience. As reported in Tables 1 and 2, the within-group differences (rows) were as large as the differences between piece-rate groups (columns). This finding emphasizes Adams' position that inputs and outcomes must be evaluated in terms of the worker's perception of them, rather than in terms of objective reality.

Worth noting in the reported row differences is the fact that the performance of the \$2.25-or-more group was only slightly different from the performance of the \$1.60–\$2.20 group—while the performance of the \$1.55-or-less group was markedly different. This fact might mean that there was a ceiling on worker pay expectations for a campus job, and once a worker's previous wage experience was high enough to bring him up to that limit, further increases in previous high wage had no additional effect.

From an applied viewpoint the within-piece-rate-group differences suggest that previous wage history should be considered in hiring new employees. Other things equal, an applicant whose previous high wage was less than the wage for the present job might be a better risk than an applicant whose previous high wage exceeded the wage for the present job.

Limitations in Generality

As always there are several considerations which limit the generality of this study and the related studies cited in this report. First, there is the time factor: that which is true for a 2-hour work period might not be true for longer work periods. Second, that which is true for a college-student population is not necessarily true for other populations. Among unskilled workers, for example, social class differences in belief value systems might lead to a different response to overpay inequity. Or, in a management population, dedication to achievement for its own sake might counteract the effects of underpay inequity. Third, in all the studies reported, each worker was isolated from all other workers. This fact is a handicap when you try to generalize to the more typical work situation where workers are involved in one or more groups. Until such time as the above considerations have been tested by additional research, caution must be exercised in drawing inferences for industrial practice. The experimental data tell only that a perceived wage inequity does influence the job performance of college students who are hired for short-term jobs which allow them to work in isolation from other workers.

REFERENCES

ADAMS, J. S. Toward an understanding of inequity. *Journal of Abnormal and Social Psychology*, 1963, 67, 422–436. (a)

ADAMS, J. S. Wage inequities productivity and work quality. *Industrial Relations*, 1963, 3, 9–16. (b)

ADAMS, J. S. Injustice in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 2. New York: Academic Press, 1965. Pp. 267–297.

ADAMS, J. S., & JACOBSEN, P. R. Effects of wage inequities on work quality. *Journal of Abnormal and Social Psychology*, 1964, **69**, 19–25.

ADAMS, J. S., & ROSENBAUM, W. E. The relationship of worker productivity to cognitive dissonance about inequities. *Journal of Applied Psychology*, 1962, **46**, 161–164.

ARROWOOD, A. J. Some effects on productivity of justified and unjustified levels of reward under public and private conditions. Unpublished doctoral dissertation, University of Minnesota, 1961.

JACQUES, E. *Equitable payment*. New York: Wiley, 1961.

LAWLER, E. E., & O'GARA, P. W. The effects of inequity produced by underpayment on work output, work quality and attitudes toward the work. Paper read at American Psychological Association, New York, September 1966.

SAMPSON, E. E. Status congruence and cognitive consistency. *Sociometry*, 1963, **26**, 146–162.

(Received February 23, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 46–49

VISUAL FACTORS AFFECTING TRANSFER OF TRAINING FROM A SIMULATED TO A REAL CONTROL SITUATION

M. HAMMERTON and A. H. TICKNER

Medical Research Council, Applied Psychology Research Unit, Cambridge, England

3 experimental groups, each of 12 Ss, were trained to control a trolley moving on a miniature railway so as to carry out an acquisition task. They were trained using (a) a TV display, (b) a cathode ray tube (CRT) with correctly scaled photographic transparency, (c) CRT with out-of-scale transparency, and (d) plain CRT. A control group of 12 Ss trained on the real trolley ab initio. These conditions produced markedly different transfer ($a > b > c > d$) both initially and with a savings measure. In 6 of 12 comparisons made, differences were highly significant ($p < .01$); in 4 they were significant ($p < .05$) and in 2 not significant. A further experimental group (12 Ss), using a TV display, and trained to carry out a pursuit task, also showed very good transfer. Conclusions are: ideal simulation must take account of both (a) adequate background texture and (b) accurate depth scaling. Absence of (a) can produce stimulus compounding; absence of (b) leads to misestimations.

A previous study (Hammerton, 1963) reported an investigation devised to examine transfer of training from simulated to real control tasks, both being relatively simple and dynamically identical. The real situation in that investigation was an acquisition task involving the control of a trolley moving upon a miniature railway at a distance from the subject (*S*) of 60 yards. The kinematically identical simulation display was presented on a cathode ray tube (CRT). It was then found that the simulator training produced very poor first trial transfer, though the recovery of skill was very rapid. It was suggested that these findings might be an instance of stimulus compounding in the transfer situation, which is known to give such an effect (Estes, 1959).

If this were so, it would be predicted that a simulator display which included a good deal of the general appearance, background, and texture of the real situation would give markedly better transfer than the plain CRT previously used. This condition could be closely approximated by using, as the training "simulator" display, the monitor screen of a closed-circuit television (TV) apparatus. However, should such a device give the very good transfer expected, the hypothesis would not stand proved; for another factor could be involved: the effects of visual scaling of depth and speed. For in the real situation there is present all the rich variety of depth and speed cues which are normally available; the TV display gives a large number of these, for example, overlap, convergence, texture, known objects, etc. (see Gibson, 1950), though imperfectly; whereas all the cues provided by a CRT relate to blips of light very close to *S*, and moving at a few centimeters per second. In order to examine the importance of this factor, it was necessary to compare transfer to the real situation from two displays, both of which gave a lot of "real world" texture, but of which one was markedly out of scale.

The experiments reported here were therefore devised to investigate the effect of these factors upon transfer. The real situation was identical with that described in Hammerton (1963). In the first experiment a TV system was used for training; in the second and third a CRT was used, modified by covering the screen with photographic transparencies of two different magnifications in the two experiments.

In considering the results of these experiments it was necessary to consider the possibility that, in the conditions where accurate background details were supplied, *Ss* had merely learned to respond to particular features of that background. For example, had *Ss* learned such things as "centralize control

as soon as the trolley passes such-and-such a landmark?" The first experiment was therefore repeated with a pursuit task. In this condition such learning would not be possible (since no particular background features would be correlated with particular trolley-target separations) and, if it were important, much poorer transfer would be expected.

METHOD

Subjects. The control group for Experiments 1–3 (data incorporated from Hammerton, 1963), the control group for Experiment 4, and each of the experimental groups consisted of 12 Royal Naval ratings, whose ages ranged from 18 to 26 years.

Apparatus and task. The real object controlled by S was a small wheeled trolley running along a straight length of railway, normal to his line of sight. The S and the experimenter (E) sat in a control cabin 60 yards from the track. The S used a thumb joystick to control the movements of the trolley, the linear velocity of which was proportional, apart from inherent lag, to the deflection of the stick. The control gain was approximately 300 seconds^{-1}, that is, a stick deflection of 1 centimeter produced a velocity of the trolley of 300 centimeters per second.

Scoring was by means of a dekatron timer, which E started with a press button, and which, in Experiments 1–3, stopped when the center of the trolley had remained continuously for 2 seconds within a particular zone of the track. The length of this zone was 6 inches, and its size and position were indicated to S by a white-painted metal target, which appeared behind the track when the dekatron was started and which fell when the dekatron stopped. Fixed to the middle of the trolley was an upright strip of black metal; when this strip was in line with the white target, the trolley was within the target zone. The trolley always started each run from a position 6 yards to the right of the target. The S's task was to move the trolley to the target as rapidly as possible, and to keep it there until the target fell. His score on each run was the time shown on the dekatron, which included the 2-second "holding time."

During training sessions in Experiment 1, Ss (referred to as Group A) observed this real situation on a TV screen. The TV equipment used was a Pye industrial closed-circuit set, using a type 2026 camera, a type 2317 control unit, and a type 2870 monitor screen. This gives a 405 line picture on a 11.5-inch × 8.5-inch screen. The camera was mounted on a tripod beside the cabin at the level of S's head. (The viewpoint of the picture was thus 1 yard to the left of S's viewpoint in transfer trials.) The screen was so placed that, at S's eye, the angular movements of the trolley were approximately correct.

In Experiments 2 and 3 the real situation was simulated on a CRT, so that, for a given control movement the angular display movement at S's eye was identical in the simulated and real cases. This was achieved by picking up the trolley's movements directly. As the trolley moved, it operated a potentiometer, which controlled the X-plane voltage of one output of a double-gun CRT. The other output was fixed as a vertical line, and the voltages were so arranged that the "trolley" spot coincided with the fixed line when the trolley was on target. The fixed line thus represented the target; it was made to appear and disappear as the target itself did. By these means the linear movements of the spot upon the CRT were in strict correspondence with those of the trolley itself, and no question of error in representation arose. The distance from starting point to target on the screen was 50 millimeters. At S's eye this subtended the same angle as the 6-yard run at 60-yard range. Thus, whatever control movement S made, the angular display movement was identical in both real and simulated situations.

During simulation sessions, the CRT was mounted (as was the TV monitor during Experiment 1) on a frame outside the control cabin, with its face against the window through which S normally looked, the rest of the window being blanked out.

The face of the CRT was covered by a photographic transparency of the view which S would have in the real situation. In Experiment 2 this was exactly to scale, so that the spots of light on the CRT appeared upon the photograph precisely where the trolley and target would appear in the transfer situation. In Experiment 3, however, the transparency was enlarged by a factor of 2, so that, though angularly correct, the spots of light appeared to refer to smaller movements against a very much closer background. The groups taking part in these experiments are referred to as Groups B and C, respectively.

In Experiment 4 the apparatus and task were altered. A 12-yard length of rail was laid parallel to the existing one and (from S's viewpoint) behind it. Along this new rail ran a trolley carrying a white-painted metal target 6-inches square. This trolley moved back and forth along its rail at a uniform velocity of 5 feet per second. (This was approximately half the maximum velocity attainable by the trolley under S's control.) Micro-switches at each end of the track reversed the movements of the target when it came to them. An angled strip of metal was attached to S's trolley so that it interposed between a lamp and a photocell which were on the target trolley when S's trolley was "on target." A short-range radio device was used to operate a relay in the scoring equipment when this happened.

The S's task was to bring his trolley from the opposite end of the target's run, to align it with the target, and to keep it aligned for 2 seconds continuously. When this had been done, the target stopped, the run ended, and S was told his score. Trolley and target were then returned to the ends opposite to those from which they had started, and the next run could begin (i.e., runs started from alternate ends of the target's range).

The group taking part in this experiment is referred to as Group X.

TABLE 1

TRANSFER MEASURES OBTAINED UNDER THE SEVERAL EXPERIMENTAL CONDITIONS

Experimental group	Training condition	ϵ	σ
A	TV	0.98	1.00
B	CRT with correct phototransparency	0.71	0.92
C	CRT with × 2 transparency	0.39	0.85
D	Plain CRT	0.24	0.77
X	TV of moving target	0.88	0.91

Procedure. This was identical in all the experiments. The *Ss* were shown the apparatus, and the nature of the task and method of scoring were carefully explained to them. The *E* demonstrated the task once, and trials began.

The *Ss* in each experimental group carried out 10 trials each day for 3 days using the TV monitor, or CRT with transparencies.

They were told the time they had taken immediately after each trial; they thus had 30 trials with knowledge of results. On the fourth day the training display was removed, and *Ss* were given 10 trials in the real situation, again with knowledge of results. The *Ss* in the control group were similarly given 10 trials each day in the real situation for 4 days.

RESULTS [1]

It is by no means easy to select the most suitable measures for transfer from simulated to real tasks, and the choice finally made must depend upon the questions being asked. Two measures of transfer are used: one is a "savings measure," which deals with the saving of training time required to reach a stable performance level in the posttransfer task compared with the time taken by the control group; the second is a "first-shot" measure, which deals with the amount of pretransfer training which is retained in the immediate posttransfer situation (Hammerton, 1966).

The savings measure used here was constructed as follows: using the Mann-Whitney *U* test (Siegel, 1956) it was found that none of the experimental groups differed significantly from the relevant control group during the second half of Day 4 ($p < .01$ in all cases). It was found that the scores of the control group for Experiments 1–3 did not differ reliably from one another, at the 5% level, after their thirteenth trial. If then for any of experimental Groups A, B, and C, successive group mean scores do not differ reliably (again at the 5% level) after the nth transfer trial, the savings measure σ was defined as:

$$\sigma = \frac{13 - n}{13}$$

The control group of Experiment 4 required 22 trials to reach a stable performance level. (Their task being a more difficult one, they naturally took longer to master it.) For Group X, therefore, σ was defined as

$$\sigma = \frac{22 - n}{22}$$

The "first-shot" measure used was internal to each group. This was ϵ, defined as

$$\epsilon = \frac{F - T}{F - L}$$

where F is the performance on the first training trial, L on the last, and T is the performance on the first transfer trial. Such an internal measure could be deceptive if the stable performances of the experimental groups differed significantly from those of the control group. As noted above, however, this was not the case here. The values of ϵ and σ for each of the experimental groups are laid out in Table 1. (The data for the fourth row, Group D, are incorporated from Hammerton, 1963.)

The entries in the ϵ column are the means of the several *Ss*' ϵ scores. The studentized

TABLE 2

SIGNIFICANCE OF DIFFERENCES (p) BETWEEN GROUPS ON THE 2 MEASURES OF TRANSFER

Groups compared	"First-shot" measure (ϵ)	Savings (σ)
A vs B	<.05	<.05
A vs C	<.01	<.01
A vs D	<.01	<.01
B vs C	<.05	*ns*
B vs D	<.01	<.01
C vs D	*ns*	<.05

[1] The authors wish to thank P. Altham for advice on statistical techniques.

range statistic (Winer, 1962) was used to examine the significance of the differences between the ϵ values of Groups A–D. The significances are shown in Table 2. Since the task of Group X was different from that of the other groups, to include their results in this statistical comparison would hardly be valid. That the values are very high will, however, be noted. As the σ values are based upon the group pooled means, significances had to be examined indirectly. This was done by fixing the posttransfer stability point for each *S*, using the one sample runs test (Siegel, 1956), and using the above method to test for differences in n and hence σ.

DISCUSSION

It should be recalled that there are three groups of factors concerned in any situation involving the transfer of a visual motor skill. These are the system dynamics, the visual display, and the other sensory features of the display, for example, kinesthetic and auditory elements. Our studies have been confined to the second of these. The good transfer values obtained by Group X strongly indicate that merely contingent details of the display were not a serious confounding factor in our results. Pains were taken to ensure that the several training and transfer situations did not differ in their dynamics, and it is difficult to find accounts of other experiments in which this was done. Even in those where it was, for example, Dougherty, Houston, and Nicklas, 1957, there was great variation in the kinesthetic properties of the training and transfer situations. Thus the present series of experiments must stand by itself as an examination of the visual factors in transfer from simulated to real control situations.

To attempt a numerical comparison of the importance of the various factors is not at present possible. It is possible, however, to assert that really high first-shot transfer requires (*a*) adequate depth-scaling cues, and (*b*) adequate background detail. It seems demonstrated that absence of (*b*) can, in fact, produce stimulus compounding in the transfer situations.

The practical significance of these findings clearly depends upon extrapsychological (especially economic) considerations. It seems that, when the initial transfer is poor, even when markedly so, the effect is transient and recovery is rapid. Therefore, if the requirement of a simulator is that it should save training time, a relatively simple visual display can be quite sufficient. The desirability of good initial transfer must be weighed against the increasing cost of realistic simulation.

REFERENCES

DOUGHERTY, D. J., HOUSTON, R. C., & NICKLAS, D. R. Transfer of training in flight procedures from selected ground training devices for aircraft. *USN Training Device Center, Technical Report*, 1957, No. 71-16-16.

ESTES, W. K. The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: A study of a science.* New York: McGraw-Hill, 1959. Pp. 380–491.

GIBSON, J. J. *The perception of the visual world.* Boston: Houghton-Mifflin, 1950.

HAMMERTON, M. Transfer of training from a simulated to a real control situation. *Journal of Experimental Psychology*, 1963, **66**, 450–453.

HAMMERTON, M. Measures for the efficiency of simulators as training devices. *Ergonomics*, 1966, **9**, in press.

SIEGEL, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received February 25, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 50–64

DEVELOPMENT OF MODERATOR VARIABLES TO ENHANCE THE PREDICTION OF MANAGERIAL EFFECTIVENESS [1]

ROBERT HOBERT

Humber, Mundie & McClary, Minneapolis

AND

MARVIN D. DUNNETTE

Psychology Department and Industrial Relations Center, University of Minnesota

Item analyses were used to develop 2 moderator variables which, on cross-validation, successfully identified managers who were over- and underpredicted by regression equations developed earlier. These moderator tests were used to identify and eliminate from the cross-validation sample 55 managers (25% of the total) classed as "unpredictable." The point-biserial correlation between predicted criterion status (above or below the criterion median) and actual criterion status for the remaining managers was .73 as compared with a point-biseral correlation of only .65 for all managers in the sample. Moreover, the degree of overlap on the actual criterion scale between managers predicted to be "high" or "low" was reduced from 38% to 28% by using the 2 moderator tests. It is concluded that these results provide further confirming evidence of the usefulness of moderator variables for enhancing the magnitude of relationships in test validation and selection research.

There are 27 years separating the publication of two frequently cited statements concerning test validation and selection research: Clark Hull's *Aptitude Testing* in 1928 and Edwin Ghiselli's *The Measurement of Occupational Aptitude* in 1955. Hull argued that validity coefficients of about .50, corresponding to a forecasting efficiency of only 13%, represented the upper limit for tests of the time, and stated that forecasting efficiencies above this were inaccessible for test batteries of that day. In his summary of both published and unpublished validity studies, Ghiselli confirmed the prophetic nature of Hull's dictum when he found that nearly all validities were in the range .30–.40 with a magnitude of .50 a distinct rarity.

Twenty-seven years with little discernible improvement in predictive accuracy provide incontrovertible proof of the necessity for changes in methodology. Thus, Ghiselli's review is probably one of the major influences leading to recent methodological assaults on the problem of low validity coefficients in test and selection research.

Attempts to update the theoretical model underlying test validation and selection research have emphasized the need to reject the oversimplification of the classic prediction model and the need for a model giving more realistic cognizance to the complexities involved in predicting human behavior. A recent model suggested originally by Guetzkow and Forehand (1961) and modified by Dunnette (1963a, 1966) is shown in Figure 1. It takes account of the

> complex interactions which may occur between predictors and various predictor combinations, different groups (or types) of individuals, different behaviors on the job, and the consequences of these behaviors relative to the goals of the organization [Dunnette, 1963a, p. 318].

This new model explicitly defines the intervening variables, that is, the individuals, job behaviors, and situations, which can affect the relationship between predictors and cri-

[1] This study was completed as a doctoral dissertation submitted to the Graduate School of the University of Minnesota by Hobert under Dunnette's direction. The authors are deeply grateful to Harry Laurent of the Standard Oil Company of New Jersey for providing the test and criterion information for the managers taking part in Standard's Early Identification of Management Potential (EIMP) study.

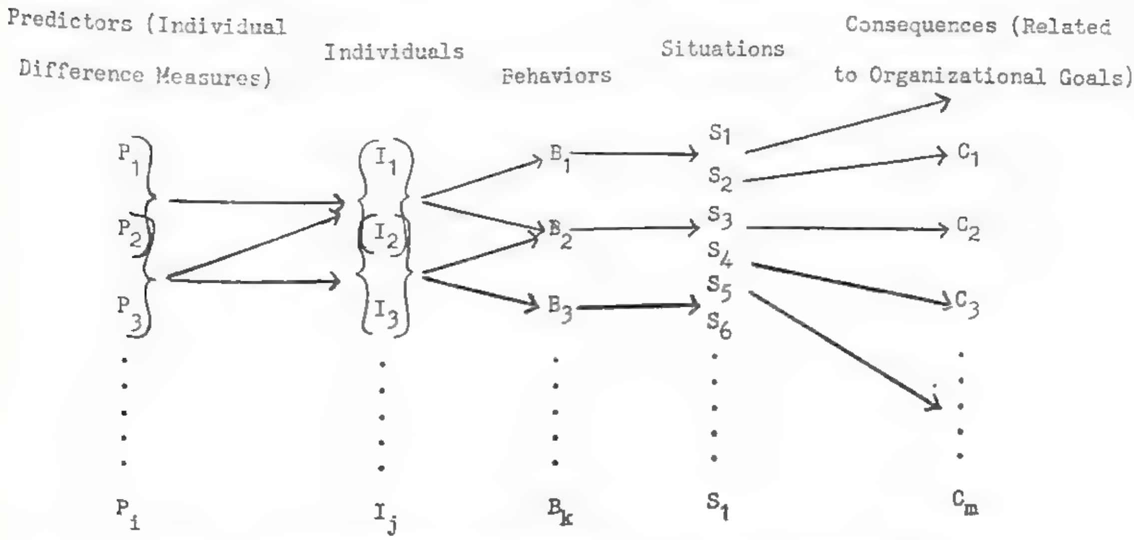


Fig. 1. A model for test validation and selection research (Dunnette, 1966).

teria. Instead of ignoring such factors, as in the classic validation model linking predictors to criteria by a simple index of relationship, the Guetzkow-Forehand-Dunnette model implies that prediction can be enhanced by determining how these intervening variables operate in given situations. Thus it dictates that validation studies need to become more specific by defining and taking account of the unique conditions appropriate for different job situations, persons, and specific behavioral outcomes. Thus, in order to take account of complex interactions between predictors and organizational consequences, it is necessary to identify relatively homogeneous subsets of predictors, individuals, job behaviors, and situations. For example, the J coefficient developed by Primoff (1955) identifies subgroups of jobs based on careful job analyses and emphasizes the differential predictability of job activities. Balma, Ghiselli, McCormick, Primoff, and Griffin (1959) also recognized the varying predictability of job activities in their discussion of *indirect* or *synthetic* validity.

Several investigators have suggested ways of identifying these homogeneous subsets of individuals or "types." Summarizing a variety of methods of assessing profile similarity, Gaier and Lee (1955) and Cronbach and Gleser (1953) conclude that these indexes are merely variants of the formula for the linear distance between two points in *n*-dimensional space. Lykken (1956) has taken issue with the psychological meaning of similarity defined in terms of linear distance, and has proposed a method of actuarial pattern analysis requiring no assumptions concerning the form of the distribution and defining similarity in psychological terms. He accomplishes this by classifying subjects (*S*s) together into cells on the basis of similar test scores, and then investigating the criterial outcomes for persons in each cell. Toops' (1959) use of biographical and test similarities for developing "ulstriths" is essentially the same as Lykken's approach, and he also advocates developing different prediction equations for each subgroup.

Another approach to the definition of "types" of individuals has involved the discovery and use of moderator variables. Although investigators have given different names to the variables they develop, for example, "population control variable" (Gaylord & Carroll, 1948), "modifier variable" (Grooms & Endler, 1960), "referent variable" (Toops, 1948, 1959), "predictability variable" (Ghiselli, 1956, 1960a, 1960b), and "moderator variable" (Banas, 1964; Saunders, 1956), these variables have in common the implication that homogeneous subgroups of individuals may be isolated which will show different patterns of validity.

The challenge to update their models and techniques thus is being met by those en-

gaged in test and selection research. In fact, the very number of new techniques proposed for handling the various factors in the prediction equation has created a new need. This new need is for research devoted to theoretical questions, to comparing the various techniques under a variety of conditions, and to refining the new techniques and explaining more precisely how they work. Such research must be undertaken now to lay the groundwork, by accumulating a body of evidence, for a future synthesis of these techniques into a workable theoretical model.

It was in response to this need that the present study, concerning the use of moderator variables in identifying homogeneous subgroups of more and less predictable individuals was undertaken.

Related Research and Statement of the Problem

In his survey of the historical development and current status of research on moderator variables, Banas (1964) proposes the following definition of a moderator variable: It is

> The general term to refer to all variables, quantitative or qualitative, which improve the usefulness of a predictor by isolating subgroups of individuals for whom a predictor or set of regression weights are especially appropriate.

That such moderators exist is proved by the number of investigators who have been able to find or develop them. What is less clear is how to find or develop these moderators.

In reviewing the literature, Banas noted the following empirical approaches to the identification of moderator variables: (*a*) An absolute-difference approach which considers deviations from a Line of Relations, where items are selected which correlate most highly with the absolute differences between the standard scores on the predictor and the standard scores on the criterion (Ghiselli, 1956, 1960a, 1960b, 1963); (*b*) A "deviate technique" in which item responses are correlated with the difference between actual and predicted criterion measures and the resulting scales are used as predictors in a multiple-regression equation (England, 1960; Neidt & Malloy, 1954); (*c*) Assessment of intraindividual variability, where variability among subscores on a test is used to determine more and less predictable subgroups of individuals (Berdie, 1961); and (*d*) Statistical infrequency of response and/or response inconsistency approaches which have been used to develop verification, validity, or carelessness moderator scales for the SVIB (Fillbeck & Callis, 1961), the Kuder Personal Preference Record (Kuder, 1956), the Minnesota Vocational Interest Inventory (Campbell & Trockman, 1963), and the MMPI (Meehl & Hathaway, 1946).

Another approach suggested by Banas is to "develop a moderator by item analysis in which the algebraic differences between the standard predictor scores and the standard criterion scores are used to select the items." Implicit in taking an algebraic difference is a notion of the importance of direction of prediction, that is, overprediction and underprediction. Whereas the absolute-difference technique divides a sample into two groups, "hits" and "misses," or as Ghiselli calls them "predictable" and "unpredictable," taking an algebraic difference results in three categories: "predictable," "overpredicted," and "underpredicted." Overprediction and underprediction have practical meaning both to the institutional decision maker and to the individual for whom a prediction is to be made. Instead of combining both subgroups into a single "unpredictable" category, the concepts of over- and underprediction pay heed to possible differences in the psychological makeup of persons whose scores exceed their criterion status (overpredicted) and those whose scores fall below their criterion status (underpredicted).

A rather coarse way of identifying over- and underpredicted persons is by examining the quadrants of the scatter diagram depicting the relationship between a test battery and the corresponding criterion scores. By dividing both the predictor and criterion scores at the median, individuals can be classified as follows: *high hits:* high predictor–high criterion; *overpredicted:* high predictor–low criterion; *low hits:* low predictor–low criterion; and, *underpredicted:* low predictor–high criterion. Each of these subgroups has a single unique combination of predictor-cri-

terion scores; thus, the quadrant analysis provides more homogeneous subgroups than the algebraic-difference technique by the subdivision of the "predictable" portion of the sample. The approach is depicted pictorially in Figure 2.

In terms of the modified model, the "Job Behavior" and "Situation" segments are not dealt with directly by any of these methods of developing moderators. However, by defining subgroups of individuals in terms of predictor-criterion scores, possible effects of job behaviors and situations may be taken into account indirectly.

There are several theoretical as well as practical implications resulting from this approach. First, since the overpredicted and underpredicted groups differ from each other both on predictor scores and criterion scores, it seems reasonable that they are two distinct groups differing in certain characteristics important in the prediction situation. This suggests that a single moderator, developed by combining most of the individuals in these two groups as in the absolute-difference approach, would probably mask these potentially important differences. Therefore, the one moderator used in the absolute-difference approach should be less effective in enhancing prediction than the two moderators implied in the quadrant-analysis technique, that is, one moderator developed on and used separately for each of the poorly predicted groups.

Second, even though the algebraic-difference technique also entails two moderators, the groups contain a much greater spread of predictor-criterion scores than is the case in the subgroups falling in the various quadrants. In addition, each of the two moderators in the quadrant analysis is developed on separate and distinct "predictable" groups, while the two moderators in the algebraic-difference approach both use the same predictable group in their development. Therefore, some dilution of the effectiveness of the moderators seems unavoidable in the algebraic-difference method.

Third, since it is generally known that failure is usually more easily predicted than success, one might reasonably expect to obtain differential validities for low-predicted

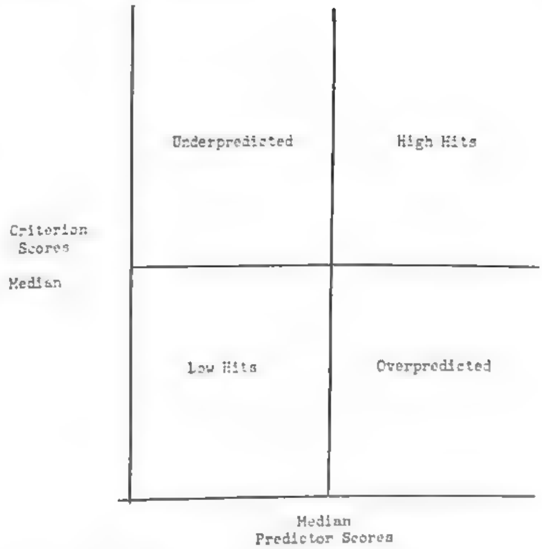


FIG. 2. General diagram of subgroups resulting from quadrant analysis. (Predictor and criterion scores are used in a general sense, the predictor score referring to the single test battery score used to predict the criterion and the criterion score referring to the single score used as a measure of success in selection or test-validation research studies.

and high-predicted subgroups. The quadrant analysis provides an opportunity to incorporate such a finding, while the algebraic- and absolute-difference techniques do not.

Fourth, by subgrouping on the basis of predictor-criterion score combinations, the quadrant-analysis approach makes possible a more complete analysis of the criterion. By comparing the quadrants having common criterion but different predictor scores, various different ways of achieving similar criterion scores may be studied. Such criterion analyses will become increasingly important as the emphasis on a single global measure of job success diminishes and the complexity of the criterion is brought under scrutiny (Dunnette, 1963a, 1963b).

Fifth, from a practical standpoint, the four subgroups in the quadrant analysis are easy to obtain, requiring none of the computations necessary in the algebraic- and absolute-difference approaches. For example, the subgroups can be identified immediately from a scatter diagram by drawing two lines intersecting at right angles through the median predictor and median criterion scores.

The above arguments seemed sufficiently compelling to warrant trying to develop two

sets of moderator variables designed to identify more and less predictable subgroups of managers in a study of the predictors of managerial effectiveness.

METHOD

Sample

The sample consisted of 443 managers of the Standard Oil Company of New Jersey and affiliated companies who had participated in a study designed to determine management potential early in an employee's career.[2] This sample is particularly appropriate for this study for several reasons. First, the sample is large enough to permit the subgroupings necessary in the quadrant-analysis approach while also assuring a reasonably large group for cross-validation. Second, the consequences of wrong decisions in management selection, both to the individual and to the corporation, are such that no possibility of improving these decisions should be left unexplored. Yet, with the exception of one study where executives were included along with factory workers and foremen (Ghiselli, 1963), no effort appears to have been made to determine whether or not moderators can be developed to identify more and less predictable subgroups of managers. Third, the Early Identification of Management Potential (EIMP) study offers a unique test of the moderator concept. Whereas most, if not all, previous moderators have been developed on predictors correlating .50 or less with the criterion, the shrunken multiple *R*s resulting from double cross-validation in the EIMP study were .71 and .70. Perhaps there exists a level of predictive validity beyond which the moderator concept will have diminished value. If so, this ought to become apparent in studying the "predictability" of the EIMP group.

The 443 managers in this sample ranged in organizational level from Chairman of the Board and President of the parent company to the second level of supervision and, in some instances, to the first or lowest level of supervision.[3] Staff specialists at organizational levels equivalent to those of the managers were also included. All participation was on a voluntary basis, and the managers were told that they were taking part in a research project in which individual results would not be used for administrative purposes.

The average age of the sample, at the time they were tested, was about 48 years ($SD = 7.5$), and they had an average of 21 years of service with the company ($SD = 9.0$). Of the 56% who were college graduates at the time they were hired, 36% also held graduate degrees. All but 3% of the sample were or had been married, and 6% were born outside the United States. Representing the top 2 or 3% of the 135,000 employees in the Standard Oil of New Jersey organization, their salaries in 1955 ranged from about $14,000 to $250,000.

The two subsamples used in the original EIMP study, Sample A consisting of 222 *S*s and Sample B with 221 respondents, were used in the present study. Sample A was used as the experimental sample on which all moderator variables were developed, and the experimental results were applied to Sample B for cross-validation.

Criterion

The criterion of management "success" used was the overall success criterion developed in the EIMP study. Although this criterion score is a single measure, the recognition given to the complex nature of the managerial job and to the various factors contributing to success in management makes this score superior to other global criterion scores typically obtained on managers. A description of the development of this criterion will indicate the composite nature of this single score.

The success criterion was developed from three main criterion measures: position level, rated managerial effectiveness, and salary history. Position level was defined roughly as the level in the organizational hierarchy to which the man had advanced. Its use as a criterion was based on the assumption that, in general, the managers who had advanced to the higher position levels had demonstrated greater managerial ability.

A ranking of managers on the basis of their effectiveness was included in recognition of the differences in effectiveness existing between managers at the same or nearly the same position level. The method used was an alternation-ranking procedure. The rankers were other managers at higher levels in the organization. Lists of individuals, containing from 6 to 30 names who were at the same general position level and in the same type of activity, were ranked by at least three higher-level managers. As some evidence of the reliability of this procedure, a test-retest reliability of .93 was obtained when 50 rankers ranked 1,284 employees twice within a few days' time.[4]

[2] The Early Identification of Management Potential (EIMP) study was begun in 1955, and the major portion of the research was completed in 1961. The research was designed to shed light on two operational problems of concern to the managers of the Standard Oil Company of New Jersey: (*a*) how to determine success in management; (*b*) how to identify employees who have the potential to be successful in management positions.

[3] This description of the sample, the criterion, and the predictor is extracted from a paper describing the original EIMP study entitled "The Identification of Management Potential," presented by Harry Laurent, Assistant Manager, Employee Relations Research Division, Standard Oil of New Jersey, at a Division 14 meeting of the American Psychological Association, St. Louis, September 1962.

[4] This reliability study was conducted independently of the major research work. The results are contained in Volume 3 of Social Science Research

Salary history, the third criterion variable, was used in the belief that the amount of salary received by an employee is based, at least in part, on how well he does his job and the value placed on that job. While recognizing the obvious limitations of salary as a single index of performance, it was still felt that salary information could add to the overall estimate of a manager's effectiveness. Three salary criteria were used: a standard score salary, averaged over a 5-year period on salary figures adjusted for the effects of inflation; an average salary for the same 5-year period adjusted for age; and a measure of the rate of salary progress during each man's career with the firm.

The overall success criterion resulted from a factor analysis of these and many other potential criterion variables. The factor accounting for the greatest variance had high loadings on the variables considered to be the primary criteria, that is, position level, rankings of effectiveness, and salary, and was thus designated the success factor. To overcome modest positive loadings on age and length of service, this success factor was rotated obliquely to yield a final overall success factor independent of age and length of service.

Using this success factor, criterion scores were calculated for each participant, and then converted to stanines. It is this stanine score which was used as the criterion in the present study.

Predictor

The predictor developed in the EIMP study, and used in the present study, was as complex as the criterion. The predictor, too, was reduced to a single score, but once again it gave due recognition to the ability of individuals to obtain this score by different configurations of test performances, as described below:

The test battery consisted of an interview, three standardized tests, and several experimental instruments requiring, in total, about 8 hours of administration time. The standardized tests were the Miller Analogies Test, Form H (Advanced Personnel Test), the Non-Verbal Reasoning Ability Test (published by Richardson, Bellows, Henry, and Company), and the Guilford-Zimmerman Temperament Survey. In addition to the 10 standard scales of the Guilford-Zimmerman Survey, experimental scoring keys were developed through item analyses of the instrument.

Another experimental instrument used as a predictor was the Individual Background Survey (IBS). It contains a series of questions covering the areas of home and family background, education, vocational planning and experience, finances, leisure-time activities, health history, and social and community relations. Its development and use reflected previous successes with this type of background information in selection research studies within the company.

A Management Judgment Test, comprised of problem situations with several possible choices of action or decision, was another experimental instrument included on the basis of previous successes using it in selecting first- and second-line supervisors. The purpose of this test was to determine the examinee's judgment insofar as managerial problems are concerned and to compare his judgment with that of other executives.

Other experimental instruments included a Survey of Management Attitudes, a Self-Performance Report, a TAT-type Picture Technique, and a Personal History Record. Since these instruments carried less weight in the prediction equation and thus were not used in the present study, a further description of them will not be undertaken.

In the EIMP study, an item analysis of this entire test battery, consisting of over 1,000 items with about 5,400 alternative responses, was run against the overall success criterion. In terms of zero-order correlations, the Individual Background Survey correlated highest with the criterion, followed by the Management Judgment Test and the experimental key on the Guilford-Zimmerman Temperament Survey.

The multiple-regression equations, calculated for each of the two subsamples in a double cross-validation design, resulted in shrunken multiple *R*s of .71 for Sample A and .70 for Sample B. Scores based on the multiple-regression equations described above constituted the predictor variable used in the present study. The predictor scores used for Sample A in the present study were, of course, developed from Sample B and cross-validated on Sample A, and vice-versa for the predictor scores used for Sample B. The modest differences in regression weights in the two multiple-regression equations made the predictors somewhat less comparable than a predictor score resulting from a single, combined multiple-regression equation would have been.

Development of Moderators

The *S*s in Sample A were divided into persons scoring above and below the median on the predictor scores. Each of these groups was then divided at the median of the overall success criterion. However, since the criterion score was expressed as a stanine score, 41 individuals obtained the median score of 5. Placement of these 41 persons was done as follows: It was reasoned that those with predictor scores below the median were underpredicted (because they were at the median on the criterion), and that those with predictor scores above the median were overpredicted (again because they were at the median on the criterion). Figure 3 shows the quadrants resulting from this reasoning and the numbers of individuals from Sample A in each quadrant.

The item pool for developing the Moderator Tests included all of the questions in the In-

Reports, New York: ESSO Standard. Copies of these reports may be obtained by writing to Harry Laurent of the Standard Oil Company of New Jersey.

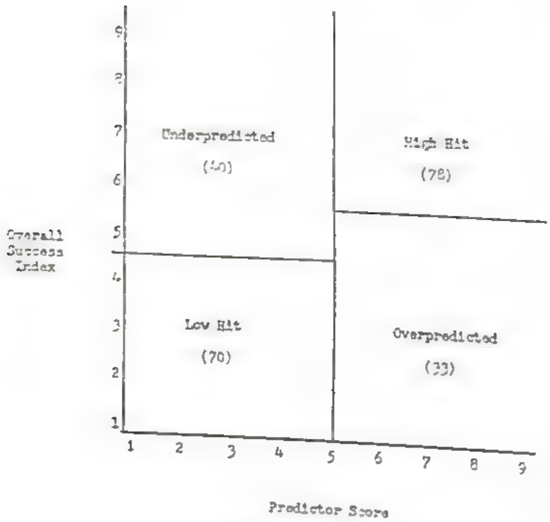


FIG. 3. Quadrants formed for Sample A and numbers of subjects in each one.

dividual Background Survey, the Guilford-Zimmerman Temperament Survey, and the Management Judgment Test, a total of 624 items. Since these instruments had the highest zero-order correlations with the criterion and also had the greatest weight in the regression equations, the items from these instruments as well as the scale scores seemed the most promising for developing tests to identify predictable and unpredictable subgroups of individuals. Thus, an item moderator test and a scale moderator test were developed for the two subgroups having low predictor scores in common, that is, the low-hit and underpredicted subgroups. Two moderator tests were also developed, independently, for the two subgroups having high predictor scores.

The percent difference statistic was used in selecting item responses for inclusion in the moderator tests. An item or scale score response was included if the percent difference between the high and low groups was significant at the 5% level in a two-tailed test. In the present case, the percent difference necessary for significance at the .05 level was roughly 20%. Unit weighting was used for the item responses included in the moderator tests, since the case for differential weighting of item responses is difficult to justify (Guilford, 1954, 1956) on the grounds of a practical increase in predictive efficiency.

The three moderator tests, item, scale, and total (i.e., the combination of weights from both item and scale differences), developed on the low predictor score subgroups, were scored so that a high score represented underprediction. For the high predictor subgroups a high moderator test score represented overprediction. Again, unit weighting was used throughout.

The method used to test the effectiveness of these moderator tests in identifying more and less predictable subgroups of individuals is implied in the quadrant-analysis approach. Since the two subgroups on which the moderator test was developed have common predictor scores but different criterion scores, the moderator test scores should correlate with the criterion scores. Therefore, if the moderator is effective, there should be a positive correlation between moderator test scores and criterion scores for the low predictor groups in the cross-validation sample, and a negative correlation for the high predictor groups. The greater the magnitude of these correlations, the more effective is the moderator.

RESULTS AND DISCUSSION

Low Predictor Groups

Table 1 presents the results of the item analysis, the ranges of scores on the moder-

TABLE 1

RESULTS OF ITEM ANALYSIS, RANGES OF SCORES ON ITEM, SCALE, AND TOTAL MODERATOR TESTS, AND MODERATOR TEST–CRITERION CORRELATION FOR ITEM, SCALE, AND TOTAL MODERATOR TESTS DEVELOPED ON THE LOW PREDICTOR GROUPS

	Item moderator test	Scale moderator test	Total moderator test
Item variable analysis			
Number of items weighted	31	16	47
Percentage of difference range	20–32%	20–47%	20–47%
Moderator test score range			
Developmental sample	7–25	0–16	7–38
Cross-validation sample	7–21	2–14	12–33
Moderator-criterion correlation			
Developmental sample[a]	.61**	.65**	.71**
Cross-validation sample[b]	.11	.31*	.27*

[a] $N = 110$.
[b] $N = 121$.
* $p < .05$.
** $p < .01$.

TABLE 2

RESULTS OF ITEM ANALYSIS, RANGE OF SCORES ON ITEM, SCALE, AND TOTAL MODERATOR TESTS, AND MODERATOR TEST–CRITERION CORRELATION FOR ITEM, SCALE, AND TOTAL MODERATOR TESTS DEVELOPED ON THE HIGH PREDICTOR GROUPS

	Item moderator test	Scale moderator test	Total moderator test
Item variable analysis			
Number used	30	12	42
Percentage of difference range	20–31%	20–44%	20–44%
Moderator test score range			
Developmental sample	7–24	1–12	9–34
Cross-validation sample	9–23	0–11	9–31
Moderator-criterion correlation			
Developmental sample[a]	−.71**	−.68**	−.77**
Cross-validation sample[b]	−.23*	−.55*	−.46**

[a] $N = 111$.
[b] $N = 99$.
* $p > .05$.
** $p > .01$.

ator tests designed to identify the underpredicted group, and correlations between moderator test scores and criterion scores for the developmental and cross-validation samples.

These results indicate that the scale moderator test and the total moderator test were effective in identifying more and less predictable subgroups of individuals having low predictor scores in common, while the item moderator test fails to "hold up" on cross-validation. Effectiveness, as used in this context, refers to the ability of the moderator to discriminate between two subgroups of individuals both of whom attained similar scores on a test battery, but different criterion scores. From a practical standpoint, this

TABLE 3

DISTRIBUTION OF SCALE MODERATOR TEST SCORES FOR UNDERPREDICTED AND LOW-HIT SUBGROUPS IN SAMPLE A AND METHOD USED TO DETERMINE OPTIMUM CUTTING SCORE

	Cumulative frequency			Cumulative percentage		Index of differentiation (Low hit %—underpredicted %)	
	Score	Low hit	Underpredicted	Low hit	Underpredicted		
	0	2	0	3	0	3	
	1	6	0	9	0	9	
	2	9	0	13	0	13	
	3	13	0	19	0	19	
	4	16	0	23	0	23	
	5	26	0	37	0	37	
	6	36	1	51	3	48	
	7	43	3	61	8	53	
	8	49	7	70	18	52	
Optimum cutting score	9	55	9	79	23	56	Index of greatest differentiation
	10	61	12	87	31	56	
	11	64	18	91	46	45	
	12	67	24	96	62	34	
	13	70	28	100	72	28	
	14	70	34	100	87	13	
	15	70	38	100	97	3	
	16	70	39	100	100	0	

TABLE 4

DISTRIBUTION OF SCALE MODERATOR TEST SCORES FOR OVERPREDICTED AND HIGH-HIT SUBGROUPS IN SAMPLE A AND METHOD USED TO DETERMINE OPTIMUM CUTTING SCORE

	Cumulative frequency			Cumulative percentage		Index of differentiation (High hit % − overpredicted %)	
	Score	High hit	Overpredicted	High hit	Overpredicted		
	1	4	0	5	0	5	
	2	15	0	19	0	19	
	3	22	0	28	0	28	
	4	36	0	46	0	46	
	5	44	1	56	3	53	
	6	52	3	66	9	57	
Optimum cutting score	7	65	7	82	21	61	Index of greatest differentiation
	8	72	14	91	42	49	
	9	77	22	98	57	31	
	10	78	27	99	82	17	
	11	79	31	100	94	6	
	12	79	33	100	100	0	

moderator test is therefore effectively identifying the individuals for whom the original prediction, based on the test battery, would have been incorrect. These results provide empirical evidence that individuals having similar predictor scores but different criterion scores do show stable and predictable differences.

High Predictor Groups

Table 2 presents results of the item analysis, the ranges in scores on the moderator tests designed to identify the overpredicted group, and the correlation between moderator test scores and criterion scores for the developmental and cross-validation samples. These results indicate that all three moderator tests are effective in identifying more and less predictable subgroups of individuals having high predictor scores in common.

The results of the analysis on the high predictor groups are therefore substantially the same as those for the low predictor groups. The major differences, a significant item moderator test correlation and substantially greater correlations for the other two moderator tests developed on the high predictor group, are merely differences in magnitude.

Effect of Moderators on Validity

The true test of a moderator is whether or not it improves the validity of a test battery. Therefore, we determined the extent to which use of these moderators would enhance the validity of the EIMP test battery.

Method. Only the two scale moderator tests were used because they were most effective in identifying their respective subgroups. A cutting score was determined, using the method described by England (1961), for each scale moderator test on *Ss* in Sample A such that the percentage of overlaps between hit groups and the over- and underpredicted groups was minimized. These optimal cutting scores were then used to eliminate individuals with high scores from Sample B (cross-validation sample), as is usually done with moderator tests. Since the resulting prediction is dichotomous, that is, either high or low criterion score, a point-biserial correlation was calculated in order to determine the overall estimate of validity.

Results and discussion. Tables 3 and 4 present the distributions of scale moderator test scores in Sample A for low and high predictor groups, respectively. In addition,

TABLE 5

DISTRIBUTION OF SCALE MODERATOR TEST SCORES IN SAMPLE B FOR LOW PREDICTOR AND HIGH PREDICTOR SUBGROUPS AND PERCENT OF SUBJECTS ELIMINATED BY CUTTING SCORES

Low predictor subgroup		High predictor subgroup	
Score	N	Score	N
0	0	0	2
1	0	1	5
2	5	2	5
3	6	3	12
4	5	4	8
5	15	5	14
6	14	6	15
7	17	7	18 Cutting score
8	9	8	8
9	15 Cutting score	9	6
10	7	10	5
11	16	11	1
12	6		
13	4		
14	2		
Total	121	Total	99
Number eliminated 35 = 29%		Number eliminated 20 = 20%	

they show the method used to determine the optimum cutting score.

Table 5 shows the distribution of scale moderator test scores in Sample B for both predictor groups as well as the appropriate cutting scores and percentage of *Ss* eliminated. Table 6 presents the calculation of the point-biserial correlation for Sample B after *Ss* scoring high on the two scale moderator tests have been eliminated.

Recognizing that the multiple correlation used in the original EIMP study is not exactly comparable to the point-biserial correlation, a point-biserial correlation was calculated for Sample B with all *Ss* included using the original regression scores to predict either a high or low criterion score as in the quadrant-analysis approach. The regression predictor scores were cut at the median for purposes of predicting high or low criterion scores. The resulting point-biserial correlation is shown in Table 7.

These results indicate that, at least in the present case, use of the moderator variables developed by the quadrant-analysis approach does enhance predictability. Thus, by eliminating only 25% of the *Ss*, the cross-valid point-biserial validity coefficient was increased from .65 to .73, and the degree of overlap between high and low criterion groups was decreased substantially—from 38% to 28%. The difference in the two validity coefficients is significant at the .05 level.

Another way of viewing these results is presented in Table 8. The percentage of accurate predictions is shown based on the regression equation alone and for the regression equation plus moderators. Two different prediction problems are presented in this table, predicting a criterion score from 1 to 5 and from 5 to 9, or in the other case, predicting a criterion score from 1 to 4 and 6 to 9. This dual prediction was a result of the modified quadrants used in the present study. Due to the standard criterion scores used, the underpredicted moderator (low predictor group) predicted *Ss* having criterion scores of 5–9, while the overpredicted moderator (high predictor group) predicted criterion scores of 1–5. Therefore, both mod-

TABLE 6

POINT-BISERIAL CORRELATION TABLE, MEAN CRITERION SCORES, AND OVERLAP STATISTICS FOR SAMPLE B WITH SUBJECTS ELIMINATED ON THE BASIS OF QUADRANT-ANALYSIS SCALE MODERATOR TEST SCORES

Actual criterion score	Predict low criterion (0) N	Predict high criterion (1) N
9	1	9
8	0	11
7	2	20
6	2	18
5	12	16
4	25	5
3	22	0
2	17	0
1	5	0
Totals	86	79
Mean criterion score	3.488	6.544
Overlap[a]	28%	
r_{pbi}	.73	

[a] The overlap figure is estimated from Tilton's (1937) Tables. The figure denotes the percentage of scores in one distribution which are matched (overlap with) scores from a second comparison distribution.

erator tests predict a criterion score of 5, and in this sense a criterion score of 5 is always considered either underpredicted or overpredicted.

The accuracy of prediction in terms of hit rate indicates again that the moderators improved on the regression equation. While the absolute improvement was small, the improvement was considerable relative to the range remaining within which to accomplish improvement.

It should be noted, however, that one effect of these moderator variables is to remove cases (as unpredictable) located toward the center of the predictor distribution. This is shown by the *SD*s of 2.81 and 3.14 obtained before and after applying the moderator variables. However, the effect is by no means the same as simply dropping cases randomly from the middle of the distribution. The 55 cases identified as "unpredictable" occur in a score range containing 188 of the original cases. Thus, the moderators identify properties other than "merely" being in the central range of the predictor distribution. Examination of the scatter diagram shows that one such property is that the 55 so-called unpredictables are cases that generally are more distant from the line of relations than the 133 cases retained in the sample. Thus, the variables are operating as moderators usually do, rather than "just" increasing the variance of the predictor set. Even so, the tendency of the moderators to delete *S*s lying in the central portion of the distribution suggests that it would have been wiser and possibly more efficient to have used tests and scales other than those already in the predictor set as a basis for developing the moderator variables.

TABLE 7

POINT-BISERIAL CORRELATION TABLE, MEAN CRITERION SCORES, AND OVERLAP STATISTICS FOR SAMPLE B USING ONLY REGRESSION PREDICTORS

Actual criterion score	Predict low criterion (0) N	Predict high criterion (1) N
9	1	9
8	0	12
7	3	20
6	9	25
5	22	25
4	33	8
3	29	0
2	18	0
1	6	0
Totals	121	99
Mean criterion score	3.760	6.303
Overlap[a]	38%	
r_{pbi}	.65	

[a] See Footnote a of Table 6.

TABLE 8

PERCENTAGE OF ACCURATE PREDICTIONS BASED ON REGRESSION PREDICTION ALONE AND REGRESSION PLUS MODERATOR PREDICTION

Criterion scores predicted	Regression alone		Regression plus moderator	
	No. predicted	Percent accurate	No. predicted	Percent accurate
1–5	121	89	86	94
5–9	99	91	79	94
Total	220	90	165	94
1–4	121	71	86	80
6–9	99	67	79	74
Total	220	69	165	77

Description of Underpredicted and Overpredicted Groups

The content of the variables and items in the moderator tests proving effective for identifying underpredicted and overpredicted individuals was examined. The purpose of this examination was not so much to determine exactly why an individual was incorrectly predicted, but rather to see if any hypotheses could be developed concerning different types of unpredictability.

Looking first at the variables, it was discovered that 10 of the 12 variables identifying the overpredicted group were also included in the moderator test for identifying the underpredicted group. However, the scoring for these variables was exactly the opposite for the two groups. Thus, on each of the 10 variables where a high score characterized the underpredicted group, it was a low score which characterized the overpredicted group,

and vice-versa. This finding in itself is of considerable consequence for it suggests that the underpredicted and overpredicted groups may not only be different, but opposite.

The content of these variables which identify both underpredicted and overpredicted individuals is not too surprising. On the E scale of the Guilford-Zimmerman Temperament Survey, the underpredicted individuals tend to score high, indicating emotional stability, composure, cheerfulness, and evenness of mood. The lower scores obtained by the overpredicted *Ss* are indicative of the opposite.

Two other keys, O and F on the Guilford-Zimmerman, characterize the overpredicted group as suspicious, hypersensitive, and getting into trouble (O), as well as having less respect for others, less acceptance of domination, and less toleration for hostility (F). These two keys did not appear in the moderator test identifying the underpredicted group.

The underpredicted group was characterized by two keys developed experimentally as having a need, as well as the ability, to dominate, and to have a good awareness of their own skills and abilities. These two keys were also used to identify the overpredicted group, but their scores indicated a lack of ability and need to dominate, as well as a lack of knowledge of their skills and abilities.

On six keys developed experimentally by running an item analysis of biographical and test items against different success factors, the underpredicted group typically had high scores while the overpredicted group had low ones. Although the item content of these keys is not known, it is further evidence that underpredicted individuals really belong to the high criterion group rather than the low criterion group predicted by their test scores. Conversely, the overpredicted person more closely resembles individuals in the low criterion group than in the high group.

In the only item moderator test which was effective, the one developed on the high predictor group, the item content indicates that overpredicted individuals tend to worry more about their work, they feel the ability to get along with others is less important, and their home life is rated less satisfactory than is true of the high-hit group. In addition, high-hit individuals are more active in public affairs, come from more successful families, and generally have had a more well-rounded education and socially complete early life.

Putting all of these variables together, the image of the underpredicted group emerges as one characterized by emotional stability, skills in interpersonal relations, self-confidence, broad perspective, self-insight, dominance, and aggressiveness. Overpredicted individuals can be characterized quite simply as lacking these qualities.

This description of the underpredicted group bears a striking resemblance to the time-honored "American success formula." That such characteristics are related to success was noted by Laurent (1962) when he found successful individuals describing themselves on the Guilford-Zimmerman "as being more stable emotionally, more objective, more friendly, and more able in their personal relationships" than the self-descriptions of less successful individuals.

The hypotheses developed, based on these characterizations of underpredicted and overpredicted groups, have a theoretical as well as a practical significance. The descriptions of the two wrongly predicted groups suggest the hypothesis that there is a common factor underlying the differences between underpredicted and overpredicted individuals. Thus the difference between underprediction and overprediction is related to quantitative differences on the same variables, that is, high and low scores, rather than to qualitatively different variables.

These descriptions further suggest that the common factor underlying these differences is made up of variables which are in some way related to the criterion or success factor. If this is the case, perhaps future moderator tests can be developed rationally, in terms of the relationships between items or variables and criteria, rather than sifting through a host of items which increase the spurious correlation and decrease the effectiveness of the moderator.

In terms of the modified prediction model, the interaction between predictors and individuals creates different organizational out-

comes for some individuals with similar overall predictor scores. The moderator tests, using a subset of these same predictors, apparently are able to specify which type of interaction will occur and how the outcome will be affected by identifying the variables (predictors) which are most likely to offset or compensate for either high or low scores on the remaining predictors. Thus we have a differential weighting of predictors for this study, applied sequentially, to improve classification by identifying the compensatory or offsetting types of predictors in the prediction situation.

IMPLICATIONS

This study gives support to the theoretical notion that the quadrant-analysis method is a feasible and effective technique for developing moderator variables. Moreover, the approach seems capable of increasing the concurrent validity of a set of "predictor" scores even when the magnitude of the unmoderated relationship already is substantial.

In an insightful analysis of the manner in which moderator variables appear to function, Lykken and Rose (1963) suggested a theoretical model which seems particularly appropriate to the quadrant analysis undertaken in this study. Their "explanation" of the moderator-variable model is stated as follows:

> The predictability of Y from X varies as a function of Z, although Z may be uncorrelated with either Y or X. One circumstance—perhaps the only one—which might lead to such an outcome would be where the "form" of the function relating Y to X differed markedly among individuals in the sample. For example, suppose $Y = aX$ for part of the sample while $Y = b/X$ for the remainder . . . if Z is the likelihood that an individual belongs to the $Y = aX$ group (i.e., if Z is a discriminant function), then we should be able to show high correlations between Y and X for persons with high Z and high correlations between Y and $1/X$ for persons with low Z [p. 142].

In the terms of this model the predictable subgroup would be the $Y = aX$ part of the sample and the unpredictable subgroup would be the $Y = b/X$ segment remaining. Although the definition of the predictable subgroup is appropriate for the function $Y = aX$, the definition of the unpredictable subgroup is not appropriately handled by a single function such as $Y = b/X$. Instead, the unpredictable group really consists of two separate groups of persons, those above and below the so-called line of relations (the line corresponding to $Y = X$). Theoretically, prediction for these two different groups would *not* be handled optimally by any single function such as $Y = b/X$. Moreover, it is unlikely that there would be a high correlation between Y and $1/X$ in *both* unpredictable groups. It would seem, therefore, that the absolute-difference method is actually inappropriate to the model described by Lykken and Rose, and that we may, in the long run find it to be an inefficient and impracticable method for developing moderator variables.

In contrast, the definition of more and less predictable subgroups by the quadrant-analysis approach is more appropriate to the Lykken-Rose model. Here, the low-hit and high-hit groups contain the persons for whom the regression line, $Y = aX$, is particularly appropriate whereas the underpredicted and overpredicted groups are better described by $Y = b/X$. However, in the quadrant approach, Z (the moderator) *does* correlate with Y. The correlation results from developing two moderator tests instead of only one. By truncating the range of predictor and criterion scores found in subgroups, the opportunity for positive and negative correlations to cancel out is reduced. Moreover, since a major difference between the low-hit and underpredicted subgroups is their criterion (or Y) scores, it is logical that a moderator test designed to identify the two subgroups will be related to several elements in the criterion. The same is true for the moderator test designed to identify the overpredicted and high-hit subgroups. Therefore, the relation between moderators and criterion is consistent with the model in this case.

The model also accounts for the fact that the overpredicted were characterized as lacking the same traits which typified the underpredicted. In fact, the model would predict just such an outcome. Since the $Y = b/X$ line theoretically extends downward from left to right, that is, from the underpredicted quadrant through the overpredicted quadrant, a high score characterizing a trait for the underpredicted individual should appear in

the profile of the overpredicted person as a low score or lack of that trait.

If the underpredicted and overpredicted individuals differ only quantitatively on the same variables, and if the same is true for the low-hit and high-hit groups, both of which are implied in the Lykken-Rose model, then it may be possible to identify the four quadrant-analysis subgroups with only one moderator test. By developing a moderator test, say, on the underpredicted and low-hit subgroups, high scores on the moderator test would indicate underprediction and low scores would indicate an accurate low prediction. This same moderator test could then be extended to the overpredicted and high-hit subgroups by simply reversing the interpretation of high and low moderator test scores. Thus, for the high-predictor subgroups, a high moderator test score would indicate an accurate high prediction and a low score would indicate overprediction.

Perhaps the most important implication of these results is the possibility of their leading to a renewed emphasis on the concepts of "overprediction" and "underprediction" as opposed to those of "overachievement" and "underachievement." As has been suggested by Thorndike (1963), the concepts of over- and underachievement imply a research stance which seems to view predictor and criterion scores as the infallible elements in the prediction situation and individual behavior as the fallible element. In contrast, the concepts of over- and underprediction imply fallibility in the instruments of prediction and direct a search for either different instruments or new combining rules to enhance their predictive effectiveness. It seems apparent that the latter stance is a more accurate portrayal of what is really happening in test validation and selection research and that it is therefore potentially more fruitful for generating research designed to improve the instruments and the statistics of prediction.

REFERENCES

BALMA, M. J., GHISELLI, E. E., MCCORMICK, E. J., PRIMOFF, E. S., & GRIFFIN, C. M. The development of processes for indirect or synthetic validation: A symposium. *Personnel Psychology*, 1959, 12, 395–400.

BANAS, P. M. An investigation of transituational moderators. Unpublished doctoral dissertation, University of Minnesota, 1964.

BERDIE, R. F. Intra-individual variability and predictability. *Educational and Psychological Measurement*, 1961, 21, 662–676.

CAMPBELL, D. P., & TROCKMAN, R. A verification scale for the Minnesota Vocational Interest Inventory. *Journal of Applied Psychology*, 1963, 47, 276–279.

CRONBACH, L. J., & GLESER, G. Assessing similarity between profiles. *Psychological Bulletin*, 1953, 50, 456–473.

DUNNETTE, M. D. A modified model for test validation and selection research. *Journal of Applied Psychology*, 1963, 47, 317–323. (a)

DUNNETTE, M. D. A note on THE criterion. *Journal of Applied Psychology*, 1963, 47, 251–254. (b)

DUNNETTE, M. D. Personnel selection and placement. Belmont, Calif.: Wadsworth, 1966.

ENGLAND, G. W. The interest factor in undergraduate engineering achievement. *Personnel Guidance*, 1960, 39, 401–405.

ENGLAND, G. W. *Weighted application blanks*. Dubuque, Iowa: Brown, 1961.

FILBECK, R., & CALLIS, R. A verification scale for the Strong Vocational Interest Blank. *Journal of Applied Psychology*, 1961, 45, 318–342.

GAYLORD, R. H., & CARROLL, J. B. A general approach to the problem of the population control variable. *American Psychologist*, 1948, 3, 310. (Abstract)

GAIER, E. L., & LEE, M. Pattern analysis: The configural approach to predictive measurement. *Psychological Bulletin*, 1953, 50, 140–148.

GHISELLI, E. E. *The measurement of occupational aptitude*. Berkeley and Los Angeles: University of California Press, 1955.

GHISELLI, E. E. Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology*, 1956, 40, 374–377.

GHISELLI, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. *Educational and Psychological Measurement*, 1960, 20, 675–684. (a)

GHISELLI, E. E. The prediction of predictability. *Educational and Psychological Measurement*, 1960, 20, 3–8. (b)

GHISELLI, E. E. Moderating effect and differential reliability and validity. *Journal of Applied Psychology*, 1963, 47, 81–86.

GROOMS, R. R., & ENDLER, N. S. The effect of anxiety on academic achievement. *Journal of Educational Psychology*, 1960, 51, 299–304.

GUETZKOW, H., & FOREHAND, G. A. A research strategy for partial knowledge useful in the selection of executives. In R. Taguiri (Ed.), *Research needs in executive selection*. Boston: Harvard Graduate School of Business Administration, 1961.

GUILFORD, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.

GUILFORD, J. P. *Statistics in psychology and education*. (3rd ed.) New York: McGraw-Hill, 1956.

HULL, C. *Aptitude testing*. Yonkers, New York: World Book, 1938.

KUDER, G. F. *Manual for the Kuder Preference Record, Form C*. Chicago: Science Research Associates, 1956.

LAURENT, H. Early identification of managers. *Management Record*, May 1962, 33–38.

LYKKEN, D. T. A method of actuarial pattern analysis. *Psychological Bulletin*, 1956, 53, 102–107.

LYKKEN, D. T., & ROSE, R. Psychological prediction from actuarial tables. *Journal of Clinical Psychology*, 1963, 2, 139–151.

MEEHL, P. E., & HATHAWAY, S. R. The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology*, 1946, 30, 525–564.

NEIDT, C. O., & MALLOY, J. P. A technique for keying items of an inventory to be added to an existing test battery. *Journal of Applied Psychology*, 1954, 38, 308–312.

PRIMOFF, E. S. Test selection by job analysis. Washington, D. C.: United States Civil Service Commission, Test Development Section, 1955.

SAUNDERS, D. R. Moderator variables in prediction. *Educational and Psychological Measurement*, 1956, 16, 209–222.

THORNDIKE, R. L. *The concepts of over- and underachievement*. New York: Bureau of Publications, Teachers College, Columbia University, 1963.

TILTON, J. W. The measurement of overlapping. *Journal of Educational Psychology*, 1937, 28, 656–662.

TOOPS, H. A. The use of addends in experimental control, social census, and man managerial research. *Psychological Bulletin*, 1948, 45, 41–74.

TOOPS, H. A. A research utopia in industrial psychology. *Personnel Psychology*, 1959, 12, 189–225.

(Received March 3, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 65–67

SELF-ESTEEM AS A MODERATOR OF THE RELATIONSHIP BETWEEN SELF-PERCEIVED ABILITIES AND VOCATIONAL CHOICE[1]

ABRAHAM K. KORMAN
New York University

This research tested the hypothesis that high-self-esteem students are more likely to choose those occupations which they perceive to require their high abilities than those with low self-esteem. The hypothesis was supported, thus emphasizing further the role of self-esteem as a moderator on the choice of occupational roles. Some implications of these findings for a closed-loop system whereby choice patterns which are a function of self-esteem also tend to reinforce the level of self-esteem were discussed.

The predictive power of "self-implementation" theory as an explanatory principle underlying the vocational choice process has received a considerable amount of support from various investigators in recent years (cf. Englander, 1960; Holland, 1963). Despite such support, however, it is also well accepted that there are a considerable number of other influences in this process, among the most important of which are social pressures from family and friends. It is also apparent that these influences may frequently be at variance with the wishes of the individual making the choice.

In an effort to reconcile these differing viewpoints into a common theoretical framework, the hypothesis has been proposed and tested elsewhere (Korman, 1966) that self-esteem operates as a moderator on the vocational choice process in that individuals high in self-esteem would seek those vocational roles which would be congruent with ones' self-perceived characteristics, whereas this would less likely be the case for those individuals with low self-esteem. This hypothesis stemmed from two major theoretical assumptions:

1. All other things being equal, individuals will engage in those behavioral roles which will maximize their sense of cognitive balance or consistency;

2. An individual's self-esteem or general evaluation of himself is part of his cognitive structure (self-esteem is defined as a persons' characteristic evaluation of himself as an individual; low self-esteem is characterized by a sense of personal inadequacy and an inability to achieve need satisfaction in the past; high self-esteem is defined by a sense of personal adequacy and a sense of having achieved need satisfaction in the past).

From these assumptions it was predicted and verified that individuals high in self-esteem would be more likely to choose those occupations which they perceived to be likely to satisfy their specific needs and to be in keeping with their self-perceived characteristics than those individuals with low self-esteem. Such choice patterns were predicted to be consistent with the high-self-esteem individual's cognition of himself as an adequate, need-satisfying individual and the lack of such cognitions in the low-self-esteem person.

The major purpose of the research to be reported here was to generalize these findings to the area of abilities by testing a number of hypotheses which postulated similar relationships between self-esteem, perceptions of the ability requirements of the chosen occupation, and perceptions of one's own abilities. In particular, the following hypotheses were tested:

Hypothesis 1. Individuals with high self-esteem are more likely to see themselves as having high abilities in those areas where their chosen occupation calls for high abilities

[1] This research was partially supported by a grant from the New York University Arts and Sciences Research Fund. Portions of this paper were read at the Eastern Psychological Association, New York, April 1966.

than are those with low self-esteem likely to see themselves as having high abilities in those areas where their chosen occupation calls for high abilities;

Hypothesis 2. For those areas where the chosen occupation is perceived to call for low abilities, the individual with low self-esteem is as likely to see himself as having high abilities as is the person with high self-esteem.

This latter hypothesis provides a control on the results of the first hypothesis in that, if supported, it would suggest that the high-self-esteem person does not describe himself in terms of a "set" to be more adequate in everything but rather as being more adequate just in those areas which are called for by the role.

METHOD

Subjects. The basic sample for the research consisted of 126 lower-division students at a large eastern private university, who had claimed to have made a fairly definitive occupational choice. Of these, 70 were male and 56 female. Each sex was analyzed separately.

For the two hypotheses, each sample was split into high- and low-self-esteem groups on the basis of approximately top one-third and bottom two-thirds, in line with previous research.

Measuring instruments. "Self-Esteem" was measured by the self-assurance scale of the Ghiselli Self-Description Inventory.[2] Evidence for the construct validity of this instrument is available in a number of publications (cf. Ghiselli, 1963).

"Self-Perceived Abilities" were measured by the Ability-Assessment Questionnaire. This is an instrument designed to indicate the extent to which the person sees himself as possessing various abilities which have been isolated in different factor-analytic studies. Examples of the abilities tapped by this 13-item questionnaire are "the ability to visualize or perceive objects in space and place them in relation to one another," "the ability to work rapidly and quickly with numbers," and "the ability to produce a large number of ideas in a short period of time." Using a rough classification, three of the items deal with numerical abilities, three with perceptual abilities, and the rest concern themselves with more verbal self-perceived skills. The test-retest item reliability of the questionnaire has a median of .81 (about 3 weeks apart).

"Occupationally Required" abilities, as perceived, were measured by the Career Description Questionnaire. This instrument is similar in format to the Ability Assessment Questionnaire, but differs in that the subject is asked to describe his chosen occupation. In addition, the items appear in a different order.

"Vocational Choice" was measured by a questionnaire procedure found to have high concurrent validity and reliability in the previous study (Korman, 1966). As before, all individuals who had received psychological counseling were eliminated as a control.

Procedures. The questionnaires were administered in at least two different sessions to the subjects during regular class meetings of introductory psychology classes.[3] In a few cases the Career Description Questionnaire and the Ability Assessment Questionnaire were administered at the same meeting since previous work by the author has found this to be a permissible procedure. However, generally, these were administered in separate sessions.

The average time between classes was about 2 weeks, with some considerably longer.

Discrepancy scores were computed between the two highest occupationally required abilities and the corresponding self-perceptions, and the two lowest occupationally required abilities and the corresponding self-perceptions. (In case of ties, either the top three and/or bottom three abilities were utilized.) Each individual's discrepancy score was computed as follows:

(*a*) For high abilities
Discrepancy =
Occupationally Required Abilities — Self-Perceived Abilities

(*b*) For low abilities
Discrepancy =
Self-Perceived Abilities — Occupationally Required Abilities

Thus, Hypothesis 1 predicted that the mean discrepancy would be less for the high-self-esteem group than for the low-self-esteem group for the "high abilities" analysis, while Hypothesis 2 predicted there would be no differences for "low abilities."

RESULTS

The results of the investigation are summarized in Table 1 and indicate that all hypotheses received support. That is, the high-self-esteem person sees himself as more likely to meet the ability requirements of his chosen occupation than does the low-self-esteem male.

A further check was made on these results as to whether there was any differential systematic tendency for the low-self-esteem individual to see himself as low in self-perceived

[2] The author wishes to thank Edwin E. Ghiselli for granting permission to use the Self-Description Inventory, undated.

[3] The author is indebted to James Kirkpatrick and Donald Davis for allowing the use of their classes in obtaining subjects.

TABLE 1

SUMMARY OF RESULTS-DISCREPANCIES BETWEEN SELF-PERCEIVED ABILITIES AND ABILITIES CALLED FOR IN CHOSEN OCCUPATION

Moderator	Hypothesis 1: High abilities		Hypothesis 2: Low abilities	
	Males	Females	Males	Females
High self-esteem				
M	.67	.72	.97	.73
SD	.67	.91	.97	.75
N	27	19	27	19
Low self-esteem				
M	1.23	1.09	1.10	1.10
SD	.46	.42	1.10	.74
N	43	37	43	43
t	3.73**	1.68*	.50	1.76[a]

Note.—All tests are one-tailed tests.
[a] This would have been significant at the .05 level in the opposite direction from that predicted for a one-tailed test.
* $p < .05$.
** $p < .01$.

ability, compared to occupational requirements when matched against high self-esteem. Utilizing a random sample of 53 cases from the overall sample, this was found not to be the case. The mean discrepancy between required ability and self-perceived ability, when summed over for all items (high, medium, and low), was exactly the same for the high- and low-self-esteem groups when carried out to two decimal points.

DISCUSSION

These results support quite strongly the notion that individuals with high self-esteem are more likely to seek out and accept the situations which seem to be in keeping with their own self-percept, that is, a "balance" situation. Hence, a person who thinks of himself/herself as being adequate and competent is more likely to wind up in these situations where he thinks he will be adequate and competent. Such situations are not, however, "balance" situations for those who feel they are inadequate and incompetent and hence appear to be less of an incentive for them. In fact, as the results for the females show here (and as trends in the previous research indicated), they may even tend to choose an occupation which calls more for their low abilities than their high.

Since one's own self-perceived abilities are related to one's actual abilities to at least a modest level (Arsenian, 1942) and since self-esteem results from one's self-perceived adequacy in given roles, an interesting possibility for a closed-loop system presents itself here in that the low-self-esteem individual is more accepting of situations where he does not think he will be adequate and where he actually will tend not to be adequate. This will lower his self-esteem even further and lead him even further to choose roles where he does not think he will be adequate. However, just the opposite would take place for the high-self-esteem person.

A similar closed-loop situation would present itself in the noncognitive areas in that low-self-esteem people would be more likely to accept social roles that are non-need-satisfying. This would lead to non-need-satisfying situations and thus a further lowering of self-esteem.

REFERENCES

ARSENIAN, S. Own estimate and objective achievement. *Journal of Educational Psychology*, 1942, 33, 291–302.

ENGLANDER, M. A psychological analysis of vocational choice: Teaching. *Journal of Counseling Psychology*, 1960, 7, 257–264.

GHISELLI, E. The validity of management traits related to occupational level. *Personnel Psychology*, 1963, 16, 109–113.

HOLLAND, J. C. Explorations of a theory of vocational choice and achievement: II. A four-year prediction study. *Psychological Reports*, 1963, 12, 547–594.

KORMAN, A. K. Self-esteem variable in vocational choice. *Journal of Applied Psychology*, 1966, 50, 479–486.

(Received March 4, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 68–73

WEIGHTING COMPONENTS OF JOB SATISFACTION[1]

ROBERT B. EWEN[2]

University of Illinois

Is it necessary to determine how important each component of the job is to the employee as well as how satisfied he is with each component? Evaluation of the effect of using importance measures to arrive at weighted overall job-satisfaction scores indicated that (a) scores on an 8-point scale of importance multiplied by satisfaction scores gave a "satisfaction-times-importance" total score that correlated .99 with the unweighted total; (b) the unweighted total was as highly correlated with independent measures of overall job satisfaction as any of the weighted totals. However, the most important component as determined by a ranking scale was more closely related to overall job satisfaction than the least important component. It was concluded that the present results indicated the necessity of empirically demonstrating the usefulness of importance measures before accepting total scores weighted by importance as superior to unweighted total scores.

Job satisfaction is usually measured by administering a questionnaire consisting of items dealing with satisfaction with various aspects of the job. The purpose of such questionnaires is to determine how satisfied employees are with regard to the various aspects of their job.

It has been suggested (e.g., Glennon, Owens, Smith, & Albright, 1960) that such a procedure overlooks an important dimension, namely, that of the importance of each component to the employee. If the dimension of importance is a meaningful one, the satisfaction scores for employees should be treated very differently, with the greatest weight or importance being given to the factor or factors which each employee considers the most important. As Strong (1958) has stated, "A way must be found to consider only those who are really satisfied or dissatisfied with each factor and to disregard those who do not really care about the factor [p. 451]."

The question arises, however, as to how to assign proper numerical weights to the various components. Research relevant to this question has not led to a clear-cut answer. Unfortunately, some job-satisfaction studies have used importance measures without determining whether or not the differential weighting provided by the importance measures provided any additional information not contained in the unweighted total scores (totals using unit weights for all components). For example, Glennon et al. (1960) stated that it is essential to subtract satisfaction scores from importance scores in order to get meaningful estimates of job satisfaction. However, the writers presented no empirical evidence in support of their idea. Larsen and Owens (1965b) used a 5-point scale to measure importance and multiplied satisfaction scores by importance scores to arrive at a weighted total. The relationship between the weighted and unweighted totals was not reported. In those cases where investigators have attempted to assess the merits of the importance measures, the results have been conflicting. On the positive side, Youngberg, Hedberg, and Baxter (1962) found that using importance and satisfaction measures together produced better results than satisfaction measures used alone. However, as the authors themselves stated, no outside criterion was available against which the methods could be validated. Froehlich and Wolins (1960) found that items low in satisfaction and high in importance best defined satisfaction as determined by a factor analysis. On the nega-

[1] This paper is based on a dissertation submitted in partial fulfillment of the requirements for the doctoral degree at the University of Illinois. The writer wishes to express his gratitude and appreciation to the members of his PhD committee. The writer also wishes to thank Abraham K. Korman and Joseph Weitz for comments and criticism. Thanks are also due to the companies participating (on an anonymous basis) in the study; to General Motors for permission to use the Faces Scale; and to Patricia C. Smith for providing the data used to establish neutral points for the JDI.

[2] Now at New York University.

tive side, Decker (1955) and Schaffer (1953) found that weighting by rankings on a 5-point scale did not significantly improve the correlation between the weighted sum of the components and a measure of overall satisfaction. Schaffer did find, however, that the most important component was more highly correlated with the overall satisfaction measure than was the least important component. Larsen and Owens (1965a) found that an importance scale failed to make a contribution in a study of the assignment of job-attitude items to subscales. Smith and Kendall (1963) examined the literature and concluded that:

> An unsolved problem when dealing with summed measures is that of the appropriate weighting for the areas involved. The methods that have been used include equal weighting, or weighting according to rated importance. . . . Whenever the effects of weighting have been evaluated, they have not improved either psychometric properties of the measures or their relationships with other, objective variables [p. 8].

In view of the conflicting evidence, the present study was designed with the purpose of evaluating the merits of different weighting procedures. Three separate procedures were adopted for this purpose:

1. Computation of correlations between totals arrived at using differential weightings and totals arrived at using equal (unit) weights for all components. The usefulness of this procedure has been pointed out by Ghiselli and Brown (1955); although their argument is phrased in terms of weightings of ratings of different aspects of job performance, the argument applies to the present situation:

> The coefficient of correlation between the final ratings when the items are equally weighted and the final ratings when they are differentially weighted provides the necessary index. If the coefficient is very high, then the weighting system adds nothing, whereas if the coefficient is moderate or low, the weighting system can be said to be contributing [pp. 124–125].

2. Computation of correlations of both the weighted and unweighted totals with measures of overall job satisfaction. This is the approach used by Decker (1955) and Schaffer (1953). If the weighted total does not yield a significantly higher correlation with an overall satisfaction measure than does the unweighted total, this would argue against the desirability of using weighting by importance for purposes of determining overall job satisfaction.

3. Tests of hypotheses. In addition to simply computing correlations between the various total scores and the overall measures, one may make tests of hypotheses which should hold if the totals weighted by the importance measures are more indicative of overall job satisfaction than the unweighted totals. It may be hypothesized that people who state that a component is important and who are dissatisfied with that component should show greater overall dissatisfaction than people who are dissatisfied with the component but who state that it is *not* important to them.

The purpose of the present study is to determine whether or not importance measures give useful information, over and above that provided by satisfaction scores alone, for purposes of estimating overall job satisfaction. If such is the case, the weighted totals should be more highly correlated with independent measures of overall satisfaction than are the unweighted totals; and hypothesized differences in overall satisfaction should occur as a function of *both* the importance of a component and the satisfaction with that component.

METHOD

Subjects

Subjects (*Ss*) were obtained from three large midwestern manufacturing companies.

Sample A consisted of 21 males employed in the personnel department of a company making electrical products. The *Ss*' mean age was 46; mean education, 16.3 years; mean company tenure, 18 years; mean job tenure, 10.4 years; mean annual pay, $10,900.

Sample B consisted of 22 males and 1 female doing various kinds of jobs (e.g., accountants, agricultural research workers) in a food manufacturing company. The *Ss*' mean age was 31; mean education, 16.3 years; mean company tenure, 4.5 years; mean job tenure, 1.7 years; mean annual pay, $10,100.

Sample C consisted of 80 males and 40 females doing various kinds of jobs (e.g., accountants, shop order writers, bar detailers, IBM equipment operators) in a steel manufacturing company. The *Ss*'

mean age was 33.5; mean education, 13 years; mean company tenure, 6 years; mean job tenure, 4 years; mean annual pay, $5,300.

The sample sizes stated above represent the final sample sizes after the elimination of careless *Ss*. The data for 1 *S* in Sample A, 1 *S* in Sample B, and 23 *Ss* in Sample C were dropped due to numerous omissions and/or incorrect methods of responding. (See Ewen, 1965, for a complete discussion of this problem.)

Methods of Data Collection

The methods of data collection varied from sample to sample. Ideally, all questionnaires should have been administered and collected at the companies by the writer. Unfortunately, this was not possible. For Sample A, the questionnaires were administered and collected by an official of the company. The *Ss* in Sample B were given the questionnaires by a company official and told to fill them out on their own time. The *Ss* were supplied with stamped, addressed envelopes, and mailed the questionnaires directly to this writer when the questionnaires had been completed. A total of 60 questionnaires were distributed and 24 (23 usable and 1 unusable) were returned, representing a rate of return of 40%. For Sample C, the questionnaires were administered and collected at the company, on company time, by the writer. Clearly, Sample C was the "best" of the three samples from the point of view of both sample size and data-collection methods.

Instruments

A comprehensive 22-page questionnaire was designed for use in this study. The time required to fill out the questionnaire was approximately 1 hour. The questionnaire consisted of three major parts:

Measure of satisfaction with the components. The Job Descriptive Index (JDI) developed at Cornell University by Patricia Cain Smith and her associates was selected as the measure of satisfaction with the components because of the extensive validation involved in its construction. Vroom (1964) has called the JDI "without doubt the most carefully constructed measure of job satisfaction in existence today [p. 100]." The JDI is an adjective checklist which measures satisfaction with five aspects of the job: the work itself, the supervisor, the co-workers, pay, and promotions. All JDI scales have reliabilities of .80 or higher.

Measures of importance. These included the following:

1. A traditional 8-point scale, where *S* evaluates the importance of each component of the job separately by checking an 8-point scale for each component.
2. A ranking scale, where *S* ranks the components in order of their importance to him.

In addition to the five components measured by the JDI, the importance measures also included the components of recognition, security, working conditions, and achievement.

Measures of overall job satisfaction. These included the following:

1. The Brayfield-Rothe Index (Brayfield & Rothe, 1951). This is a questionnaire consisting of 18 items dealing with the worker's feelings toward his job as a whole. The reliability of the Index is .87. Some evidence favoring the validity of this Index was reported in the 1951 paper.
2. The General Motors Faces Scale (Kunin, 1955). This is a one-item graphic scale. Five faces are presented, varying from a large smile to a large frown. The *S* is asked to check the face which most closely represents the way he feels about his job as a whole. This particular scale has not been validated, but faces scales for particular job-satisfaction dimensions have previously shown good convergent and discriminant validity (Locke, Smith, Kendall, Hulin, & Miller, 1964).

Procedure

The following methods of estimating overall job satisfaction were used:

1. Unit weights (weighting all components equally). Since importance measures were not used in this method, the total score obtained by this method was called the unweighted total.
2. The traditional 8-point scale. These importance scores and the corresponding satisfaction scores were multiplied together, yielding a "satisfaction-times-importance" weighted total.
3. First-ranked component. The satisfaction with the component ranked first in importance was taken as the estimate of overall job satisfaction.
4. First three ranked components. The satisfaction with the components ranked first, second, and third in importance was averaged to yield an estimate of overall job satisfaction.

Two additional estimates of overall job satisfaction were also included:

5. The employee's single highest satisfaction score, regardless of the importance of the component involved.
6. The employee's single lowest satisfaction score.

As a preliminary step, the neutral point for each JDI scale was calculated. This was necessary because the third method, involving the tests of various hypotheses, required that *Ss* be divided into groups which were satisfied, neutral, and dissatisfied with each component. Data were obtained from two groups of *Ss* ($N = 123$ and $N = 130$). The *Ss* consisted of male and female employees performing various duties in a large company. The data consisted of JDI scores and Face Scale scores for each component. A linear regression analysis was carried out for each component. This was done separately for each group. The neutral point on any JDI scale was taken to be the point on the regression line that corresponded to a Faces Scale score of 3.0

(neutral face). The results of the two groups were in close agreement.

Results

The first analysis involved the correlations between the weighted and the unweighted totals. These correlations are shown in Table 1.[3]

For Sample C, the scores of 19 *Ss* were discarded from the ranking scale analysis. Of these *Ss*, 18 ranked variables not measured by the JDI (e.g., security) as first in importance, and 1 filled out the ranking scale incorrectly.

The most striking finding occurred in the case of the traditional 8-point scale of importance. The correlations between the totals weighted by this method and the unweighted total were approximately .99 for all three samples. Thus, the traditional 8-point scale clearly did not add any new information to that obtained from satisfaction scores alone.

The other weighted totals were also highly correlated with the unweighted total. These correlations, however, were sufficiently low so that the variables could have quite different correlations with other variables. Since there is no definitive rule as to how high a coefficient should be before it can be concluded that the weighting methods are not adding any additional information, the reader will have to decide for himself whether or not the coefficients shown in Table 1 are sufficiently high to cast doubts on the usefulness of the importance measures.

The second analysis involved the correlations between the various total scores and the measures of overall satisfaction. These correlations, together with the intercorrelations between the overall satisfaction measures, are shown in Table 2.

Table 2 indicates that none of the importance measures yielded weighted totals which were significantly more highly correlated with the Brayfield-Rothe Index and the Faces Scale than the unweighted total. This would, of course, argue against the inclusion of im-

[3] A table showing all of the intercorrelations between the total scores may be obtained by writing to the author.

TABLE 1

Correlations between the Various Weighted Total Scores and the Unweighted Total

Weighted total	Correlation with unweighted total		
	Sample A (N = 21)	Sample B (N = 23)	Sample C (N = 120)
8-point scale	.98	.99	.99
#1 rank	.69	.63	.70[a]
#1–3 rank	.84	.71	.87[a]
"Most satisfied"	.82	.59	.81
"Least satisfied"	.85	.88	.88

[a] These correlations are based on an N of 101. See text for explanation.

TABLE 2

Correlations between the Various Weighted Total Scores and the Brayfield–Rothe Index and the Faces Scale

Totals	Sample A (N = 21)		Sample B (N = 23)		Sample C (N = 120)	
	BR	F	BR	F	BR	F
Unweighted	.73	.74	.50	.70	.66	.55
Total weighted by 8-point scale	.75	.77	.48	.68	.66	.56
Component ranked #1	.49	.54	.59	.63	.61[a]	.50[a]
Components ranked #1–3	.59	.57	.40	.51	.64[a]	.57[a]
Component ranked least important (among those measured by JDI)	.44	.24	.33	.41	.34[a]	.18[a]
"Most satisfied component"	.59	.44	.00	.24	.58	.47
"Least satisfied component"	.55	.61	.53	.75	.53	.46
Correlations between BR and F	.74		.88		.62	

[a] These correlations are based on an N of 101. See text for explanation.

Journal of Applied Psychology
1967, Vol. 51, No. 1, 74–77

RELATIONSHIP AMONG SUPERVISORS' INTEGRATION, SATISFACTION, AND ACCEPTANCE OF A TECHNOLOGICAL CHANGE

L. W. GRUENFELD AND F. F. FOLTMAN
New York State School of Industrial and Labor Relations, Cornell University

This study investigated how the integration and satisfaction of supervisors with management affected their acceptance of a technological change. Attitude questionnaires were administered to 40 first-line supervisors to measure attitude toward the change and several dimensions of integration and satisfaction. The results showed that supervisors who are relatively more integrated with the management group, more satisfied with management, and relatively high in job satisfaction are more likely to accept a management-initiated technological change.

Previous studies (Bass, 1960; Cartwright, 1952; Cartwright & Zander, 1962) suggest that individuals are more likely to accept change when they are part of the group that intends to initiate the change. Technological changes are critical events for lower-level supervisors because they are often not fully accepted members of management, and since these changes may further aggravate their problems of integration with the management group they tend to resist them. The integration problems of supervisors have been identified (Bass, 1960), but the effects of their lack of integration on their acceptance of technological change have not yet been studied. The major hypothesis of this study is that lack of integration of supervisors with the management group contributed to the failure of a technological change.

In addition, the level of satisfaction of the supervisor is likely to influence his willingness to accept a change. However, satisfaction and integration may not be independent of each other. Supervisors who are promoted from the rank and file (as was the case in this study) and who are unable to transfer their loyalty to management are more likely to be dissatisfied (Bass, 1965). Previous research shows that job satisfaction consists of many dimensions. Three of these dimensions which seemed to be relevant in this case—satisfaction with immediate superior, satisfaction with management, and job satisfaction—were selected in order to assess their differential contribution to the acceptance of the change and their relationship to integration.

Nature of the Change

Several months prior to this study the management of a steel foundry introduced an electronic data-processing system (EDP). This new system did not affect the manufacturing technology directly, but was an administrative device primarily designed to process information for accounting, scheduling, inventory control, and payroll purposes. The effectiveness of the system depended primarily on the information which supervisors provided and the accuracy of the records they kept. Not only were the supervisors the direct recipients of the change but the success of the system was directly related to their willingness to accept the additional responsibilities it required. The change failed and the EDP system was discontinued because the information supplied by the supervisors was unreliable. Interviews with supervisors also revealed that they were unwilling to assume the additional tasks required by the system.

Method

Setting and Subjects

The operations of the foundry in which this study was conducted are devoted primarily to the production of heavy castings. The operations are performed by approximately 1,200 blue-collar workers, 40 supervisors, and 3 department superintendents. The management group consists of a general manager and his immediate staff (controller, chief engineer, assistant general manager, personnel manager, etc.).

The population studied was 40 first-line manufacturing supervisors. Without exception, these supervisors have been promoted from within the work force. All have at one time been members of the

union and many of them maintain social relationships with hourly paid employees. The average supervisor has held his position for 10 years. Due to the flat structure of the organization, opportunities for promotion are extremely limited. In addition, there are ethnic and educational differences between the management group and the supervisors.

Variables

The *dependent variable* was selected in accordance with the major research interest of this study to relate integration and satisfaction of supervisors to their acceptance of a technological change. Therefore, a summated rating scale was constructed to measure the supervisor's attitude toward this change, installment of the EDP system. The items constituting this scale dealt with the supervisor's acceptance, and understanding, of the system as well as the benefits and pressures generated by it for him.

Two sets of *independent variables* were used. The first, consisting of two variables, was designed to assess the supervisor's integration with the management group. The second, consisting of three variables, was designed to assess various dimensions of supervisory satisfaction.

The integration variables were measured by the following summated rating scales:

Integration with Management: The items of this scale dealt with the extent to which the supervisor felt a part of management and the stand he took when there were conflicts of interest between management and his subordinates.

The Grapevine: These items dealt with the extent to which the supervisor relied on informal rather than formal channels of communications.

The satisfaction variables were measured by the following summated rating scales:

Satisfaction with Immediate Superior: The items of this scale dealt with the extent to which the supervisor was satisfied with his superintendent's technical, human relations, and administrative skills, his superintendent's upward influence, downward assistance, and responsibility for his own decisions.

Satisfaction with Management: These items dealt with the extent to which the supervisor was satisfied with the administrative, human relations, and technical skills of the management group.

TABLE 1

SPLIT-HALF RELIABILITY COEFFICIENTS FOR ALL SCALES USED IN THE STUDY

Variable	r_{tt}
Attitude toward EDP	.51
Grapevine	.81
Integration with management	.79
Satisfaction—immediate superior	.88
Satisfaction—management	.68
Job satisfaction	.87

TABLE 2

OBSERVED CORRELATIONS BETWEEN THE CRITERION VARIABLE AND THE INTEGRATION VARIABLES

Variable	1	2	3
1. Integration with management	1.00	−.34*	.31*
2. Grapevine		1.00	−.25
3. Attitude toward EDP			1.00

Note.—$N = 40$.
* $p < .05$ (two-tailed test).

Job Satisfaction: These items dealt with the supervisor's satisfaction with his job, pay, promotion, and progress in the company.

Reliability of Scales

The internal consistency reliability coefficients for all scales are shown in Table 1. It should be noted that they are all statistically significant ($p < .01$) and of sufficient magnitude for the analysis of group data. The above scales were administered 8 months prior to the discontinuation of the EDP system. In order to preserve anonymity respondents were not asked for any identifying information on the questionnaire.

RESULTS

Relationship between Integration and Attitude toward EDP

As hypothesized, the results show that supervisors are more likely to accept the EDP system if they are integrated with the management group.

The observed correlations between the criterion variable and the integration variables are shown in Table 2. The "grapevine" variable related significantly—inversely—with integration with management: those supervisors who report that they receive most of their information from "the grapevine" are less likely to be integrated with the management group.

The correlation between the EDP variable and the "grapevine" variable is inverse, as predicted, although it is not significant. The overall picture that emerges is quite clear: those supervisors who are relatively well integrated are more likely to view the introduction of the EDP system as beneficial.

Relationship between Satisfaction and Attitude toward EDP

As expected, attitude toward EDP is significantly related to satisfaction with both

TABLE 3

OBSERVED CORRELATIONS BETWEEN THE CRITERION VARIABLE AND THE SATISFACTION VARIABLES

Variable	1	2	3	4
1. Satisfaction—immediate superior	1.00	.21	.43**	.11
2. Satisfaction—management		1.00	.14	.36*
3. Job satisfaction			1.00	.37*
4. Attitude toward EDP				1.00

Note.—$N = 40$.
* $p < .05$ (two-tailed test).
** $p < .01$ (two-tailed test).

management and job. Satisfaction with immediate superior is significantly related to job satisfaction. However, an unexpected but important result is the lack of a significant relationship between satisfaction with immediate superior and attitude toward EDP. Moreover, two seemingly related variables, satisfaction with immediate superior and satisfaction with management, are not significantly related. These results are shown in Table 3.

Relationship between Satisfaction and Integration

Table 4 shows that some of the integration and satisfaction variables are not independent of each other. The pattern that emerges is quite consistent. Integration with management correlates significantly with all three satisfaction variables, and satisfaction with management correlates significantly with the two integration variables. Without exception, the significant relationships focus on the management group.

TABLE 4

OBSERVED CORRELATIONS AMONG SATISFACTION AND INTEGRATION VARIABLES

Variable	Integration with management	Grapevine
Satisfaction—immediate superior	.37*	−.18
Satisfaction—management	.42*	−.33*
Job satisfaction	.56**	−.16

* $p < .05$ (two-tailed test).
** $p < .01$ (two-tailed test).

DISCUSSION

The results of this study indicate that supervisors are more inclined to accept a technological change if they are well integrated with, and identify with, management. In addition, this study shows that dissatisfaction is related to lack of integration and also contributes to the unwillingness of supervisors to accept a management-initiated technological change.

Another noteworthy finding is that the supervisor's satisfaction with his immediate superior was unrelated to his willingness to accept the change. However, his satisfaction with management contributed substantially to his acceptance of the EDP system. Job satisfaction, which reflects satisfaction with management policies regarding pay, promotion, and working conditions, is also a critical determinant of the supervisor's attitude toward change. These findings suggest that the locus of resistance to change is not always with the workers and that supervisors are not immune to resisting a management-initiated change. A technological change that centralizes decision making, as did the one in this study, and in fact eliminates an entire level of supervision (in this case the superintendents), restructures the relationship between the lower-level supervisor and the management group. Under these conditions, the top management takes the place of immediate supervision at least in the view of the lower-level supervisors. Thus the introduction of the EDP system increased the saliency of top management as a source of dissatisfaction for the supervisors.

The above results are supported by interviews and observations made by the authors in this plant during the attempted installation of the EDP system and subsequent to the discontinuation of the system. These observations revealed that the integration of the supervisors with the management group prior to the technological change was unsatisfactory from the supervisors' point of view. Upward mobility was blocked, and the few opportunities for promotion that were available went primarily to individuals who were not members of the supervisory group. Supervisors maintained strong social ties with subordinates and under conditions of stress it was not un-

common for them to return to the bargaining unit. The union contract guaranteed both their right to return and the maintenance of their seniority. Supervisors also complained about inequities in pay. Due to overtime work they were handing out checks to operators, that were substantially higher than their own. Supervisors also reported that their grievances and suggestions did not receive sufficient consideration by the management group.

It was into these conditions that the management introduced the technological change. This new system was meant in part to enlarge the sphere of managerial responsibility of the supervisors and to integrate them more closely with the management group; contrary to the expectations of the supervisors the new procedures actually increased the direct intervention and pressure for production by the management and resulted in the bypassing of the superintendents. The supervisors were now in direct contact with management, but this condition, instead of facilitating integration, increased dissension between the two groups. The supervisors experienced the new relationship as close and punitive supervision. The technological change, therefore, further aggravated the problem of integration of the supervisors and resulted in their unwillingness to accept the change.

In summary, under conditions in which a technological change results in the flattening of organizational structure, lower-level supervisors will be more inclined to accept the change if they have a sense of belongingness to and satisfaction with *top* management.

REFERENCES

BASS, B. *Leadership, psychology and organizational behavior.* New York: Harper, 1960.

BASS, B. *Organizational psychology.* Boston: Allyn & Bacon, 1965.

CARTWRIGHT, D. Achieving change in people, the group dynamics approach. In L. Cook (Ed.), *Toward better human relations.* Detroit: Wayne University Press, 1952.

CARTWRIGHT, D., & ZANDER, S. (Ed.) *Group dynamics: Research and theory.* (2nd ed.) Evanston, Ill.: Row, Peterson, 1962.

(Received March 15, 1966)

Manuscripts Accepted for Publication in the *Journal of Applied Psychology*

A Factorial Study of Some Psychological, Vocational Interest, and Mental Ability Variables as Predictors of Success in Dental School: Martin K. Chen*, Dale Podshadley, and John G. Shrock: Public Health Service, Dental Health Center, 14th Avenue and Lake Street, San Francisco, California 94118.

Self and Other Semantic Concepts in Relation to Choice of a Vocation: Richard A. Hunt*: Department of Psychology, Southern Methodist University, Dallas, Texas 75222.

Factors in College Attendance: Arthur A. Dole* and John M. Digman: Department of Psychology, 2430 Campus Road, University of Hawaii, Honolulu, Hawaii 96822.

SVIB Key Length: Dissident Data: Norman M. Abrahams*: Department of the Navy, USN Personnel Research Act. (Code 521) San Diego, California 92152.

Characteristic Pace as Determined by the Use of a Tracking Treadmill: Gary L. Holmgren* and George S. Harker: Texas Instruments, Inc., P. O. Box 5474, Dallas, Texas 75222.

Alienation, Environmental Characteristics, and Worker Responses: Milton R. Blood* and Charles L. Hulin: Department of Psychology, University of Illinois, Urbana, Illinois 61801.

Determinants of Work Attitudes among Negroes: Robert Bloom and John R. Barry*: Department of Psychology, University of Georgia, Athens, Georgia 30601.

Multidimensional Scaling Analysis of Decision Strategies in Threat Evaluation: Joseph W. Rigney* and Charles H. DeBow: Electronics Personnel Research Group, University of Southern California, University Park, Los Angeles, California 90007.

Affect Level, Capillary Pulse Pressure, and Response Latency: Bruce O. Bergum* and Donald J. Lehr: Xerox, Research and Engineering Center, P. O. Box 1540, Building 14, Rochester, New York 14603.

* Asterisk indicates author for whom the address is supplied.

(Continued on page 88)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 78–81

TAT CORRELATES OF EXECUTIVE PERFORMANCE

PEARSON C. CUMMIN [1]

Harvard Business School

The TAT was used to measure n Achievement, n Affiliation, n Power, n Autonomy, n Aggression, and n Deference in 2 groups of Ss. The 1st group consisted of more successful business executives, the 2nd of less successful ones. It was hypothesized, largely from the work of Henry and McClelland, that the successful group would have significantly higher scores than the less successful group in n Achievement, n Power, and n Autonomy; whereas the less successful executives would have higher scores in n Affiliation, n Aggression, and n Deference. The successful group had significantly higher scores in n Achievement and n Power than the unsuccessful ones. The implications of these findings were discussed.

The most comprehensive delineation of the executive personality to date has been achieved by William Henry (1949). Although the individual personality is shaped to some extent by the internalization of the executive role, Henry maintains that the process of executive selection tends to favor those individuals "whose personality structure is most readily adaptable to this particular role [p. 286]." Although Henry's meticulous description of the executive personality afforded no empirical evidence, it suggested a pattern of needs which seemed to motivate the executive and which would lend themselves to empirical validation.

In *The Achieving Society,* McClelland (1961) found that high n Achievement scores (as measured by the TAT) were significantly related to "entrepreneurial" success in several different countries. McClelland defines an entrepreneurial occupation as one in which the individual has (*a*) "more responsibility for *initiating* decisions," (*b*) "more *individual responsibility* for decisions and their effects," (*c*) "more *objective feedback* of accurate data indicating the success of his decisions," and (*d*) "a job which *entails more risk* and challenge in that there is more chance of a serious wrong decision being observed [pp. 2–3]." He concludes that n Achievement and entrepreneurial success are strongly associated and that when an association is not found between these variables, then one or more of the conditions defining the entrepreneurial role have not been met.

The present study proposed to extend McClelland's findings relating n Achievement to "entrepreneurial" success. Executives who were rated as more or less successful were compared in six motives. n Affiliation, n Power, n Autonomy, n Aggression, and n Deference, as well as n Achievement, were examined. These additional motives were selected from Murray's (1943) list of needs as those which best fitted Henry's description of the executive personality.

The hypothesis tested in this study was that more successful executives would score high in n Achievement, n Power, and n Autonomy, whereas less successful executives would have high scores in n Affiliation, n Aggression, and n Deference.

The choice of the Achievement motive as one of the factors to be used in differentiating successful from less successful executives was based primarily on three considerations. First, much of McClelland's (1953, 1963) research has been centered around this motive, and he has shown that there is a strong association between n Achievement and entreprenurial success. Second, one of Henry's major characterizations of the successful business executive was "achievement desires." Third, n Achievement had already been adapted into a scoring manual by McClelland,

[1] This paper is based in part upon an honors thesis submitted to the Department of Social Relations at Harvard University in partial fulfillment of the requirements for the degree of Bachelor of Arts with Honors. The author wishes to express his appreciation to his advisor, Douglas H. Powell, for his assistance and encouragement in completing this research.

which made the question of its reliability far less doubtful.

One of Henry's characterizations of the successful business executive involved the nature of his interpersonal relations. Successful executives were described as being detached and impersonal toward their subordinates, and as having broken all emotional ties with their parents. Whyte (1956) also describes the successful executive as a man who is concerned with "getting ahead" and not just "fitting in." It was on the basis of these two considerations that the hypothesis concerning the relationship of n Affiliation to success was formulated.

The Power motive was chosen primarily because of Henry's findings that successful executives are constantly struggling for increased responsibility. Whyte refers to this mobility drive as the "executive neurosis"; but whether the successful executive enjoys moving steadily up the power hierarchy of the organization or whether he does it out of a fear of being replaced from below, the result is the same.

There has been no published research that has studied the relationship of the three remaining motives to business success. With respect to Autonomy, Whyte suggested that, while unsuccessful executives have a feeling of personal attachment toward the organization and its policies, successful executives tend to be more "out for themselves." They are not tied to the organization and will not hesitate to go elsewhere if they feel their prospects of advancement are too restricted with their current firm. This theory was generally supported by Hagen's (1959) findings that major executives tend to have achieved their success by moving diagonally from one firm to another, or even from one occupational field to another; while minor executives, who will never rise above a certain level, have tended to progress slowly and vertically up through the organization.

Henry states that a successful executive is an "active, striving, *aggressive* person," but that his aggressions are "channeled into work or struggles for status and prestige [p. 289]." Primarily on the basis of this statement, it was hypothesized that the successful executives would be low in n Aggression, whereas the unsuccessful executives would be high.

The theory on which the choice of the Deference motive was based is largely the result of an intuitive deduction on the part of the author, although there is some basis for it in Whyte's study. Whyte states that one of the major drives of the potentially unsuccessful executive is to become a part of the organization, whereas the potentially successful executive, with his well-defined self-identity, chooses to remain emotionally disconnected with his firm.

METHOD

Sample

The subject (*S*) sample and the experimental conditions are described in detail in Moment and Zaleznick's (1963) *Role Development and Interpersonal Competence*. Fifty-two business men from Boston and the surrounding area were divided into four groups. The first group consisted of engineers and technical supervisors and represented the middle level of organization hierarchy. The second group was more heterogeneous and was made up of top executives as well as men in middle management.

The *S*s in the third and fourth groups were secured a year later. The members of Group 3 were selected so as to fall in a relatively narrow age range, whereas Group 4 comprised a wide range of ages and two women. In general, however, all four groups may be said to fall within Roe's (1956) "organization-type" occupational grouping. The median age of the sample was 35. Median salary ranged from $7,500 to $11,000; and the median educational level was "some graduate work." These data were obtained through a detailed questionnaire which each *S* was required to complete.

Each *S* was also administered a written form of a specially constructed Thematic Apperception Test (TAT). These consisted of four pictures showing situations in which business men are obviously involved. The instructions given on this test were similar to the ones employed in the TAT.

Experimental Procedure

A modified version of McClelland's scoring manuals for n Achievement, n Affiliation, and n Power was used to test the presence or absence and strength of each of these three motives. The description and scoring manual for n Autonomy, n Aggression, and n Deference were devised by the present author.

Both Imagery and Need receieved a +1 as they do in McClelland's scoring system. The rationale behind this was that Imagery indicated the presence of a motive while Imagery and Need (Need was not scored without scoring Imagery) indicated an active concern over that motive. Contrary to Mc-

Clelland's system, *Ss* were not given a −1 for Unrelated Imagery, and the distinction between Doubtful and Definite Imagery was not made. Because of the small number of TAT pictures ($N = 4$) and the relatively large number of motives under consideration, it was considered impractical to score a −1 for the lack of related Imagery in any one motive. All motives not scored received an automatic 0.

The third category into which each of the six motives was subdivided was created by the author. Before devising the scoring manuals, 20 of the 48 protocols were read in order to obtain a general idea of the nature of some of the stories. It was discovered at this time that a significant number of the stories apparently adopted a somewhat ambivalent attitude toward the motive Imagery and Need which they expressed. This ambivalence often took the form of punishment, either by guilt or by other external forces, for the expression of a particular motive. Ambivalence was scored with a −1.

Each of the 48 TAT protocols was then scored independently by the author and another rater. The reliability percentages for four cards with six motives per card which were scored for each *S* ranged from 78% to 97% agreement. The reliability percentages for each motive on all 192 cards (4 per *S*) ranged from 82% to 98%. The scores of the two raters were only considered in agreement when identical for each motive on each card, and the above degree of reliability was thus considered adequate for the present study. Having scored each protocol independently, the two raters then met to discuss scoring disagreements and arrived at a consensus.

Classification of the Raw Data

The sample was divided into two groups of more and less successful executives. The distinction was made by dividing *Ss* into six age groups. The annual income of each individual was listed under the appropriate age group. The median salary was then taken for each of the six groups. All *Ss* at, or above, the median salary in their group were labeled as more successful. Those below the median were considered to be less successful.

Next, the total score of each *S* was computed for each need. The *Ss* were then divided again into their respective age groups, and the mean score of each need was computed. All scores at, or above, the mean were considered high for that particular motive; all scores below the mean were considered low.

Testing the Hypothesis

Because the hypothesis predicts the direction of influence of each motive on greater or lesser success in business, a one-tailed test of significance was employed in most cases. The chi-square technique in 2 × 2 contingency tables was used to test each specific hypothesis.

RESULTS

The results are shown in Table 1. The first hypothesis was that more successful executives would have high scores in n Achievement, whereas less successful executives would have low scores on this motive. The *p* value was significant beyond the .01 level of confidence.

The second hypothesis was that more successful executives would have low scores in n Affiliation, whereas less successful executives would have high scores. The direction of the results was opposite from that which was predicted. It is interesting to note that had the hypothesis predicted that successful executives would have ***high*** scores instead of low for n Affiliation, the value of *p* would have been .15.

The third hypothesis was that more successful executives would have high scores in n Power, whereas less successful executives would have low scores. The *p* value was significant beyond the .01 level of confidence.

TABLE 1

RELATIONSHIP OF CAREER SUCCESS TO TAT MOTIVE

Motive	χ^2	p	Hi Motive Hi Success	Hi Motive Lo Success	Lo Motive Hi Success	Lo Motive Lo Success
n Achievement	5.58	.01	19	5	10	14
n Affiliation	1.26	.30[a]	15	6	14	13
n Power	6.21	.01	18	4	11	15
n Autonomy	0.51	.25	15	7	14	12
n Aggression	0.11	.80[a]	7	3	22	16
n Deference	0.01	.50	11	8	18	11

[a] *p* value is two-tailed because results were in the opposite direction from that which was predicted.

The final three hypotheses relating success in business to n Autonomy, n Aggression, and n Deference were not supported.

DISCUSSION

On the two motives (n Achievement and n Power) which differentiated more successful from less successful business executives, the findings concerning n Achievement are probably the most significant in that they support previous research.

McClelland's research indicated a strong association between n Achievement and entrepreneurial success. The term "entrepreneurial," however, may not be specific to the business community at large. It takes into account only a small portion of businessmen. The significance of the results of the present study is that n Achievement was shown to be related to success in a much more diversified sample of the business population. The results also gave empirical validation to Henry's characterization of successful business executives as men with "achievement desires." They indicated, however, the limited extent to which Henry's characterization of successful business executives bears up under rigorous empirical validation.

The successful validation of the hypothesis concerning the power motive was considered significant because of the lack of any previous research connecting it with success in business. A chi-square showed no significant correlation between n Power and n Achievement. The findings provide empirical support for the observations of Henry and Whyte that successful executives are constantly struggling for increased responsibility. The positive findings related to power suggest that achievement striving alone is not the sole criteria for differentiating more from less successful executives. There is also a definite desire, on the part of successful executives, for increased responsibility and control within the organizational hierarchy.

The overall picture, then, of the successful executive is one of an individual who is determined to maintain a high standard of excellence in his work, and to assume greater responsibilities and more control over his environment as he advances within the organization.

These findings suggest that the analysis of motives by means of the TAT has promise for executive selection. However, at this stage, the practical use of this technique is limited. Further research supporting these findings on different populations is needed before these methods can be validly employed in industry.

REFERENCES

HAGEN, D. H. Family atmosphere and other childhood patterns as precursors of career interests. Unpublished doctoral dissertation, Harvard University, 1959.

HENRY, W. E. The business executive: The psychodynamics of a social role. *American Journal of Sociology*, 1949, **54**, 286–291.

MCCLELLAND, D. C. *The achieving society*. New York: Van Nostrand, 1961.

MCCLELLAND, D. C. n Achievement and entrepreneurship: A longitudinal study. Harvard University, 1963. (Mimeo)

MCCLELLAND, D. C., ATKINSON, J. W., CLARK, R., & LOWELL, E. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.

MOMENT, D., & ZALEZNICK, A. Role development and interpersonal competence. Boston: Harvard University Press, 1963.

MURRAY, H. A. *Thematic Apperception Test manual*. Cambridge: Harvard University Press, 1943.

ROE, A. *The psychology of occupations*. New York: Wiley, 1956.

WHYTE, W. H. *The organization man*. New York: Anchor, 1956.

(Received March 22, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 1, 82–88

MULTIPLE-REGRESSION ANALYSIS OF A PAIRED-CHOICE DIVISION-OF-TIME INVENTORY IN RELATION TO GRADE-POINT AVERAGE[1]

PURNELL H. BENSON

Consumer and Personnel Studies, Inc., Madison, New Jersey

76 college Ss reported how they would divide amounts of time per wk. (7 hr. and 21 hr.) between pairs of activities formed in all possible pairs from a list of 7 activities. Marginal utility functions for the activities were fitted to the 42 paired divisions of time for each student. The parameters of these functions were then correlated with grade-point average (GPA) in a combined regression for the 76 Ss. The correlation of the variables with GPA is increased from .28 for College Entrance Examination Board (CEEB) verbal and mathematical scores alone to .45 for activities data plus aptitude scores.

This paper reports analysis made of a fresh type of questionnaire material for recording choices concerning use of time. The underlying research aim was to estimate academic achievement from the reported choices of amounts of time to be spent upon designated activities.

METHOD

The activities and their descriptions were:

1. Study on your own: serious reading or studying beyond assignments given in courses.
2. Sports participation: team sports or anything you do for the sake of exercise.
3. Watch entertainment: looking at plays or sports, whether on television, at games, or going to movies or theater.
4. Leisure time reading: magazines, newspapers, fiction, or any other light reading.
5. Recreation with others: dating, parties, visiting or entertaining friends, but not merely staying in someone else's home and not group activities whose purpose is sports or watching entertainment.
6. Altruistic activities: volunteer social service, civic or political activities, or participating in a philanthropic school activity, but excluding work responsibilities or other categories listed here.
7. Religious activities: attending church, Sunday school, devotional reading, or other religious exercises. Activities in behalf of a church belong under the heading of altruistic activities.

The instruction to the respondent was: "If you had just 7 hours (or 21 hours) a week to spend on the following pair of activities, how would you divide your time between them?"

The 7 activities were presented in the questionnaire in the 21 possible pairs. The amounts of time to be divided by the respondent between activities in these pairs were 7 hours per week and 21 hours per week. For convenience in selecting times, a double box scale ranging by hourly steps from 0 to 7 and 7 to 0 (0 to 21 and 21 to 0) was printed next to each pair.

RATIONALE FOR FITTING UTILITY FUNCTIONS TO THE DATA

Marginal utility theory (Samuelson, 1953) provides a basis for fitting utility functions to the data for divisions of time between activities in the pairs. The usual experience of the consumer is one of decreasing increments in satisfaction from added amounts of goods consumed. With each added amount of television watched, less satisfaction is obtained, until no more TV is watched that evening or that week. The function which relates the size of the last or marginal increment in satisfaction to the amount already consumed is the marginal utility function.

In the analysis of buying behavior, it is assumed that consumers buy quantities of each commodity up to the point where the satisfaction received from the last unit of money spent upon the commodity is equal for each commodity. If the satisfaction from the last dollar spent per week upon meat is greater than the satisfaction from the last dollar spent per week upon vegetables, then the consumer will spend more money on meat and less on vegetables in order to increase her total satisfaction.

[1] Project supported by National Institutes of Health Grant No. MH-10539-01 administered by Drew University. The counsel of H. O. Gulliksen is acknowledged. Findings in this paper are the responsibility of the author.

The use made of marginal utility functions in economics is to predict how fixed amounts of disposable income are divided by the consumer between opportunities for satisfaction. At the point of division, the sum of the amounts of money spent equals the total money available to the consumer, and the marginal utilities are equal. Solution is made from these algebraic relationships for the amounts of money spent by the consumer on alternative goods.

Figure 1 represents the decreasing marginal utility lines for increasing amounts of time spent on seven categories of activity by a college student used for illustration. This figure shows graphically the marginal utility system of equations, whether for division of time or money between available opportunities. Considering the system from the standpoint of consumer buying, let the marginal utility functions for n goods be given by:

$$\begin{aligned}
u_1 &= a_1 + b_1x_1,\\
u_2 &= a_2 + b_2x_2,\\
\cdots &\quad \cdots\\
u_i &= a_i + b_ix_i,\\
\cdots &\quad \cdots\\
u_n &= a_n + b_nx_n.
\end{aligned}$$

The u's are the marginal utilities, the a's are the intercept constants for the marginal utility lines, the b's are the slope constants, and the x's are the amounts of money (or time) spent on the n opportunities for consumption. The equation of restriction specifies that the sum of the x's equals a fixed constant k.

$$x_1 + x_2 + \cdots + x_i + \cdots + x_n = k.$$

Since the marginal utilities are equal to each other, the system includes $(n + 1)$ equations with $(n + 1)$ unknowns. Solution is made for u and for the x's which specify the amounts of money (or time) spent by the consumer on the different kinds of goods.

Suppose the x's are known. Is it possible with this same system to solve for the constants which define the marginal utility functions? It is if the origin and stretch of the system are fixed arbitrarily. This can be done by designating one of the a's and one of the b's. The x's must be obtained from two or more fixed amounts of money (or time) which are divided by the consumer, in order to create enough data to solve for the remaining slope and intercept constants.

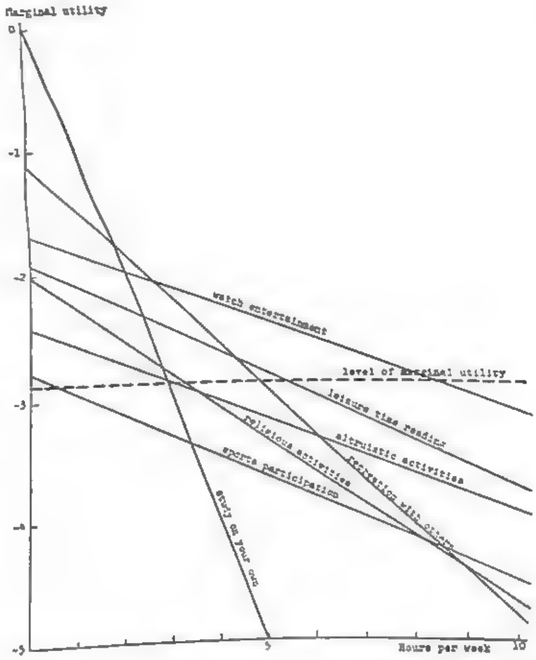


FIG. 1. Marginal utility profile lines for seven activities for case illustrated.

Change of origin or change of scale for the system of marginal utility functions does not affect the divisions of time or money between alternatives. Marginal utility functions can be used to predict consumer divisions of time or money, irrespective of the transformation applied to the zero point or stretch for the system. This circumstance makes possible an empirical determination of marginal utility functions which has apparently gone unnoticed by economists. This determination would be useful in economics for establishing demand functions from consumer purchase data.

The foregoing background of marginal utility principles indicates what is assumed in the research model for fitting functions to reported divisions of time between activities. Without prior measurement of the utilities of the activities, their function can be postulated and solution made for its parameters from the data for divisions of time. The underlying function is for the total utility of all of the activities. This is the function whose partial derivatives yield the marginal utility lines.

Let it be assumed that this function can be approximated by a parabolic surface within the region for the divisions of time. With t_i the amount of time spent upon the ith activity and U the total utility resulting from the n activities, we have, allowing for interaction terms for the dependence of activities upon each other:

$$U = a_1t_1 + \cdots a_it_i + \cdots a_nt_n + \frac{b_1t_1^2}{2} + \cdots \frac{b_it_i^2}{2} + \cdots \frac{b_nt_n^2}{2} + c_{12}t_1t_2 + \cdots c_{1i}t_1t_i + \cdots c_{1n}t_1t_n + \cdots.$$

The partial derivatives u_i of this function are linear marginal utility lines with simple interaction terms.

$$u_i = a_i + b_it_i + c_{i1}t_1 + c_{i2}t_2 + \cdots c_{in}t_n.$$

With known values for the t_i's, it is desired to define the slope line constants a_i, b_i, c_{i1}, $c_{i2}, \cdots c_{in}$. Constants are sought which minimize the errors in equating the marginal utilities. The expression minimized for divisions of time t_{ijk} and t_{jik} between activities i and j paired from the n activities and for m amounts of time k to be divided is (Benson, 1959):

$$\sum_{i=1}^{i=n} \sum_{\substack{j=1 \\ j \neq i}}^{j=n} \sum_{k=k_1}^{k=k_m} [(a_i + b_it_{ijk} + c_{ij}t_{jik}) - (a_j + b_jt_{jik} + c_{ij}t_{ijk})]^2,$$

subject to restrictions described below.

The differentiation of this expression gives a system of equations which can be cast in the pattern of a multiple-regression solution, provided certain conditions are imposed. The origin of the system, its scale magnitude, and $(n-1)$ of the interaction terms must be arbitrarily assigned. These conditions can be met by letting $a_1 = 0$, $b_1 = -1$, and the c_{1i}'s $= 0$. Furthermore, the number of parameters remaining must not exceed the number of observations of divisions of time, that is m $n\ (n-1)/2$. Examples of lines of entries for multiple regression are given in Table 1.

TABLE 1

PATTERN OF ENTRIES FOR MULTIPLE-REGRESSION ANALYSIS OF MARGINAL UTILITY FUNCTIONS

Dependent variable	Slope coefficients		Interaction coefficients	Intercept coefficients	
b_1	b_i	b_j	c_{ij}	a_i	a_j
$-t_{1ik}$	t_{i1k}	0	0	1	0
0	t_{ijk}	$-t_{jik}$	$t_{jik} - t_{ijk}$	1	-1

ILLUSTRATION OF FITTING MARGINAL UTILITY FUNCTIONS

Data for 1 of the 76 students furnishing replies to the paired-choice division-of-time inventory and for whom grade-point average (GPA) and College Entrance Examination Board (CEEB) verbal and mathematical aptitude test scores are available are given in Table 2. Included is information from the student concerning how he divided his time per week between all seven activities. These amounts of time are needed for definition of the interaction terms c_{ij}, since the amounts of time already spent upon the activities not in the pair divided must be known. When the respondent was asked to divide time between activities in the pair, he was asked to suppose that he spends his usual amounts of time upon the remaining activities not in the pair.

The lines of entries for the multiple-regression solution used are given in Table 3. The slope coefficient for the study interest was taken equal to -1 and was treated as the dependent variable. Personal judgment was used to restrict the number of interaction terms to 7, and the total parameters to 20 with 42 lines of observed amounts of time. The lines of entries were repeated in the table of input with signs reversed to eliminate sign bias in the differencing of the marginal utilities. The exploratory nature of the present study did not appear to justify a more complicated form of least-squares solution than that which was followed.

The slopes, interaction terms, intercepts, and standard errors of coefficients obtained from the multiple-regression solution for the case illustrated are given in Table 4. As a check upon the procedure, the utility function was used to estimate the way in which the student would divide the amount of time he spent upon all seven activities between the seven activities. For the case under consideration, the actual and estimated amounts of time were, in order of the activities previously listed: (1) 2, 2.9; (2) 1, .6; (3) 8, 8.4; (4) 6, 5.3; (5) 5, 4.8; (6) 3, 2.9; (7) 3, 3.2. For the entire group of 76 students, the median error in estimating amounts of time spent per week was 2.9 hours.

ESTIMATION OF GRADE-POINT AVERAGE

The constants of the marginal utility functions for each case were next tested as predictive data for estimating GPA, using all 76 cases in a single mul-

TABLE 2

DIVISIONS OF TIME PER WEEK REPORTED BY A STUDENT BETWEEN PAIRS OF ACTIVITIES

Amount of time to be divided	Amounts of time for activities paired						
	Study on your own	Sports participation	Watch entertainment	Leisure time reading	Recreation with others	Altruistic activities	Religious activities
7 hours:	—	2	5	—	—	—	—
	—	—	—	5	—	—	2
	3	—	—	—	—	4	—
	—	1	—	—	6	—	—
	—	—	4	3	—	—	—
	3	—	—	—	—	—	4
	—	—	—	—	3	4	—
	—	5	—	2	—	—	—
	1	—	6	—	—	—	—
	—	—	—	—	4	—	3
	—	2	—	—	—	5	—
	2	—	—	5	—	—	—
	—	—	5	—	2	—	—
	—	—	—	—	—	4	3
	4	3	—	—	—	—	—
	—	—	—	4	3	—	—
	—	—	4	—	—	3	—
	—	2	—	—	—	—	5
	1	—	—	—	6	—	—
	—	—	—	4	—	3	—
	—	—	3	—	—	—	4
21 hours:	—	6	15	—	—	—	—
	—	—	—	12	—	—	9
	5	—	—	—	—	16	—
	—	5	—	—	16	—	—
	—	—	11	10	—	—	—
	6	—	—	—	—	—	15
	—	—	—	—	8	13	—
	—	5	—	16	—	—	—
	4	—	17	—	—	—	—
	—	—	—	—	8	—	13
	—	5	—	—	—	16	—
	7	—	—	14	—	—	—
	—	—	11	—	10	—	—
	—	—	—	—	—	10	11
	6	15	—	—	—	—	—
	—	—	—	15	6	—	—
	—	—	13	—	—	8	—
	—	6	—	—	—	—	15
	7	—	—	—	14	—	—
	—	—	—	13	—	8	—
	—	—	12	—	—	—	9
28 hours reported spent on all 7 activities:	2	1	8	6	5	3	3

tiple regression. The marginal utility constants were readjusted so that in each case the sum of the slopes of the more popular activities (excluding studying on one's own) was equal to unity. This seemed to provide better basis for comparability of the marginal utility constants from case to case.

Multiple regressions were performed with different combinations of variables as estimators to ascertain the relative contributions of the groups of variables. The coefficients of correlation obtained are given in Table 5. These have not been adjusted for degrees of freedom, since their relative magnitudes can be

TABLE 3

LINES OF ENTRIES FOR MULTIPLE-REGRESSION FIT OF MARGINAL UTILITY FUNCTIONS

Dependent variable	Slope coefficients						Interaction coefficients							Intercept coefficients					
b_1	b_2	b_3	b_4	b_5	b_6	b_7	c_{25}	c_{27}	c_{35}	c_{45}	c_{46}	c_{57}	c_{67}	a_2	a_3	a_4	a_5	a_6	a_7
−3 −3 1 2 4 −1 −5 −6 4 7 6 −7	2 1 5 −2 −3 −2 6 5 5 −5 −15 −6	−5 4 −6 5 4 −3 −15 11 −17 11 13 −12	5 −3 −2 −5 4 4 12 −10 −16 −14 15 13	−6 3 4 −2 −3 6 −16 8 8 −10 −6 14	4 −4 5 4 −3 −3 16 −13 16 10 −8 −8	−2 4 −3 −3 5 4 −9 15 −13 −11 15 9	5 5 1 5 1 −5 −1 −5 −1 −5 1 5 11 1 5 1 −5 −1 −5 −1 −5 1	3 −1 3 1 3 −1 −3 −1 −3 −3 1 3 −1 3 1 3 −1 −3 −1 −3 −9 1	−5 −8 5 8 −5 8 −3 −8 5 8 −5 −5 −8 5 8 −5 8 −1 −8 5 8 −5	5 −6 −5 6 −5 6 −5 −6 −1 6 5 5 −6 −5 6 −5 6 −5 −6 −9 6 5	5 −3 1 3 5 −3 5 −3 −5 3 −5 5 −3 5 3 5 −3 5 −3 −5 3 −5	−5 −3 5 3 −1 −3 −5 −3 5 3 5 −5 −3 5 3 5 −3 −5 −3 5 3 5	−3 3 3 −3 −3 3 −1 −3 3 −3 3 −3 3 3 −3 −3 3 1 −3 3 −3 3	1 1 1 −1 −1 −1 1 1 1 −1 −1 −1	−1 1 −1 1 1 −1 −1 1 −1 1 1 −1	1 −1 −1 −1 1 1 1 −1 −1 −1 1 1	−1 1 1 −1 −1 1 −1 1 1 −1 −1 1	1 −1 1 1 −1 −1 1 −1 1 1 −1 −1	−1 1 −1 −1 1 1 −1 1 −1 −1 1 1

more precisely compared prior to such adjustment. The correlation coefficient of .72 for the largest regression with 31 independent variables would be reduced to .45 after adjustment for degrees of freedom.

EVALUATION OF RESULTS

While the level of correlation achieved with marginal utility constants as predictors was modest in this initial study, some indications permit optimism for further development.

1. The contribution of CEEB verbal and CEEB mathematical scores was somewhat lower than customary, suggesting that GPA used as the dependent variable was not a particularly good measure of scholastic achievement for these 76 students. The opportunity for estimating scholastic achievement with these data does not seem promising.

2. The various groups of marginal utility constants each separately contributed approximately the same explanation of variability in GPA as did verbal or mathematical aptitude. Close comparison of small coefficients of correlation which have different degrees of freedom is not possible since small coefficients of correlation cannot be accurately adjusted to obtain unbiased estimates of their magnitudes.

3. Considerable improvement in the questionnaire material seems possible. In this exploratory effort, only two amounts of time were given for respondents to divide, and only seven activities, excluding the more typical activity of studying assigned work, were given in the pairs for choices of amounts of time.

4. Instead of single measures of degree of interest which are provided by usual inven-

TABLE 4

Parameters Fitted for Marginal Utility Functions

Activity	Symbol for regression coefficient	Standard deviation of entry	Regression coefficient	Standard error of regression coefficient	Computed T value
1. Study on your own	b_1	2.459	(−1.0000[a])		
2. Sports participation	b_2	9.797	−.1731	.0607	2.852
3. Watch entertainment	b_3	5.368	−.1409	.0595	2.367
4. Leisure time reading	b_4	5.344	−.1792	.0582	3.081
5. Recreation with others	b_5	4.461	−.3659	.0932	3.926
6. Altruistic activities	b_6	4.909	−.1493	.0615	2.428
7. Religious activities	b_7	4.862	−.2650	.0795	3.333
Interaction of 2 and 5	c_{25}		−.4295	.2004	2.143
Interaction of 2 and 7	c_{27}		1.1023	.4917	2.242
Interaction of 3 and 5	c_{35}		−.6882	.5456	1.262
Interaction of 4 and 5	c_{45}		−.1684	.1356	1.242
Interaction of 5 and 6	c_{56}		−.0697	.1464	.476
Interaction of 5 and 7	c_{57}		−.1441	.1646	.876
Interaction of 6 and 7	c_{67}		−.6120	.7166	.854
Intercept for 2	a_2		−4.2023	1.7101	2.458
Intercept for 3	a_3		1.5997	2.8574	.560
Intercept for 4	a_4		−1.1214	1.0235	1.096
Intercept for 5	a_5		6.4779	4.5123	1.436
Intercept for 6	a_6		−.1689	2.2632	.075
Intercept for 7	a_7		−.1860	2.8025	.066
Intercept for 1	a_0		(.0000[b])		

[a] Value of −1.0000 was assigned as the dependent variable.
[b] Value of .0000 was assigned as the intercept for "study on your own."

tories, the marginal utility inventory provides profile lines showing how apparent desire for an activity changes with the expenditure of added amounts of time. If the profile line has a high intercept and a small slope, this indicates strong and sustained interest. If the profile line has a steep slope, this manifests rapid decline in interest. A profile line with a low intercept and a falling slope would manifest the least interest of all.

TABLE 5

Multiple Correlations of Marginal Utility (MU) Variables with Grade-Point Average (GPA)

Verbal CEEB	Math CEEB	7 amounts of time per week	6 intercepts of MU functions	7 slopes of MU functions	7 interaction terms of MU functions	SE of MU fits	$(SE)^2$ of MU fits	No. independent variables	Correlation with GPA[a]
X								1	.27
	X							1	.15
X	X							2	.28
		X						7	.33
			X					6	.25
				X				7	.39
					X			7	.38
						X	X	2	.24
			X	X	X	X	X	22	.56
		X	X	X	X	X	X	29	.67
X	X	X	X	X	X	X	X	31	.72

[a] Not adjusted for degrees of freedom. $N = 76$ cases.

The accompanying Figure 1 shows the profile lines for the seven activities in the case which was illustrated. These are lines which show the decreasing interest with each added hour per week given to the activity. In the terminology of economics, these are marginal utility lines. The intercepts were calculated on the assumption that usual amounts of time are spent per week upon the remaining activities. The usual amounts are the amounts estimated from the utility functions, rather than the reported amounts. The horizontal line shows the level of utility experienced at the last or marginal increment of time spent upon each of the activities. These details of motivational structure may assume future importance in clinical studies of academic achievement or in student guidance work.

5. The standard error of estimate of the individual functional fit is one of the predictors of achievement. The respondent who tells a consistent story about his activities raises the rating made of his future performance. This circumstance may induce respondents to report their interest more realistically since they will increase their prediction score. This feature may be useful in gaining more valid information as a basis for student counseling, as well as for predictive work.

REFERENCES

BENSON, P. H. Defining marginal utility functions by means of discriminant function analysis. *American Psychologist*, 1959, 14, 398. (Abstract)

SAMUELSON, P. A. *Foundations of economic analysis*. Cambridge: Harvard University Press, 1953.

(Received March 23, 1966)

(Continued from page 77)

Relationship between Family Backgrounds and Work Values: Frank T. Paine*, Donald R. Deutsch, and Rodney A. Smith: Department of Business Administration, College of Business and Public Administration, University of Maryland, College Park, Maryland 20742.

Motivational Effects of Knowledge of Results: The Influence of Goal-Setting: Edwin A. Locke*: American Institutes for Research, 8555 Sixteenth Street, Silver Spring, Maryland 20910.

The Herzberg Theory: A Critique and Reformulation: Carl A. Lindsay*, Edmond Marks, and Leon Gorlow: Room 206 Grange, Office of Student Affairs, The Pennsylvania State University, University Park, Pennsylvania 16802.

Auditory S-R Compatibility: The Effect of an Irrelevant Cue on Information Processing: J. Richard Simon* and Alan P. Rudell: Department of Psychology, The University of Iowa, Iowa City, Iowa 52240.

Feedback in a Complex Multiman-Machine System: Irvin L. Goldstein*, Jack F. Southard, and David A. Schum: Department of Psychology, University of Maryland, College Park, Maryland 20740.

Influence of Interpolated Periods of Activity and Inactivity upon the Vigilance Decrement: William Bevan*, Lloyd L. Avant, and Henry G. Lankford: Office of the Vice President and Provost, The Johns Hopkins University, Baltimore, Maryland 21218.

journal of applied psychology

Vol. 51, No. 2 April 1967

ON THE ASSESSMENT OF POTABILITY [1]

RONALD C. DILLEHAY
Texas Christian University

WILLIAM H. BRUVOLD
University of California, Berkeley

AND JACOB P. SIEGEL
University of Pittsburgh

Potability, the acceptance of water for drinking purposes, may be conceptualized and studied through the assessment of consumer attitudes toward domestic water. 3 attitude scales were developed and tested in 6 community studies utilizing the hypothesis that potability is a function of the total dissolved solids (TDS) in drinking water. Reliabilities of the scales were found to be .82 and higher. The hypothesis was strongly supported by the data: mean attitude scores were found to be inversely related to the TDS in community drinking water. A theory of consumer attitudes toward taste in water was briefly outlined, and the recommendation of limiting standards for TDS in domestic water was discussed.

There is an urgent need for the development of assessment procedures that can be used to objectively determine the potability of water intended for domestic use (Ongerth, Bruvold, & Knutson, 1964). Potability refers to the acceptability of water for drinking purposes. By definition, then, it must be gauged through human response to water. The questions arise of how potability can be operationally defined and assessed, and what physical characteristics of water are associated with it. Two research strategies suggest themselves as means to attack these problems. One is the traditional approach of the psychophysical methods for scaling the attributes of objects. The other approach involves social-psychological research that deals with potability of existing community water supplies. Both strategies have been developed in the program of research of which the present study is a part.

Because the social-psychological approach to the study of potability necessarily involves an assessment of consumer reaction to the taste of the community water, it falls within the domain of attitude research. In general terms the concept of attitude refers to beliefs or feelings focused on some psychological object. Taking the attitude object to be the community water, appraisal of beliefs and feelings about it, or the attribution of evaluative characteristics to it, constitutes a method of measuring potability. This report describes the construction and evaluation of three scales designed to assess consumer attitudes toward domestic water.

It presently appears that a major determinant of the potability of water is its dissolved mineral content. Minerals commonly found in water used for domestic supply include calcium, magnesium, and sodium cations in combination with bicarbonate, carbonate, chloride, or sulfate anions. These minerals can produce taste, but not odor, in water. Mineral content in water is conveniently measured in milligrams per liter, and the total mineral content in water may be expressed as milligrams per liter of total dissolved solids (TDS).

[1] The research reported in the paper was supported in part by a General Research Support Grant from the California State Department of Public Health, and in part by Contract PH 86-64-77 with the Division of Water Supply and Pollution Control of the United States Public Health Service.

TABLE 1

SCALE VALUES AND Q VALUES FOR ITEMS SELECTED FOR THE GENERAL ATTITUDE SCALE (GAS)

Scale value	Q	Item
1.17	.66	I would refuse this water even if I was starving and thirsty.
1.43	1.86	If I had a chance I would move out of this area; I am just sick and tired of this good for nothing water.
1.80	1.35	I have never experienced water as bad as our present supply.
2.04	1.69	I have a number of complaints about our water; I think the quality of our water is miserable. The water is very bad out here.
2.15	1.80	I would not advise anyone to use our present water.
2.69	1.29	The less I use our local water the better I like it.
3.18	1.80	I just can't seem to get used to our water; it still is unpleasant to me.
3.53	1.58	The water out here is bad, not terrible but bad just the same; but you learn how to live with it.
3.63	1.55	Our water creates too many problems, something ought to be done about it.
4.03	1.64	The water supply in our area is not too good; in fact we have had some problems with it in our home.
4.46	1.56	We could use a better supply of water around here; the water we now receive is not as good as it should be.
4.53	1.53	Our water would be all right if it weren't for a few disagreeable things connected with it.
4.75	1.35	Our water is all right, but I would rather use a different source.
5.24	2.00	The local water supply takes a while to get used to; besides this, the water is all right.
5.80	1.43	Although the quality of our water is nothing to rave about, I can't point to anything really bad about it.
5.89	.84	The water in our area does not affect me as being very good or very bad.
6.41	1.50	I really can't criticize our local water supply. The water is not bad and I have no complaints to make about it.
6.71	1.60	It's just ordinary water; there is nothing wrong with it at all.
7.25	1.75	Having considered all the things about our water, I think the water which we receive is pretty fair.
7.92	1.88	Our water is quite adequate; I use it all the time and like it just fine.
8.77	1.57	I use our water a lot and enjoy it.
9.04	1.67	From what I can see our water is very good.
9.54	1.83	Water is the greatest bargain we can get; it is a pleasure to use our present water; I have no problems with it.
10.00	1.31	It is a real pleasure to get such good water.
10.33	1.34	The quality of our water is perfect in almost every way.

It is conjectured that potability is inversely related to the total amount of dissolved mineral content in water (TDS). The present research was designed to provide a test of this notion and to evaluate the reliability of the attitude scales developed. Specifically, the research was designed to test the hypothesis that positivity of measured attitude toward the community water and amount of TDS in the water are inversely related. The specific prediction makes possible an evaluation of the notion regarding potability and mineral content and also the construct validity of the attitude scales.

METHOD

Construction of the Scales

Thurstone judgment procedures were used to construct three attitude scales (Edwards, 1957). One scale was designed to be a measure of general attitude toward the water, not specific to taste. The general attitude scale (GAS) was developed from 73 statements, which were constructed from letters of complaint about water supplies and from responses to an open-ended question administered in a survey by the California State Department of Public Health. The statements were judged by 44 subjects (*Ss*). Some *Ss* were students from the University of California fulfilling a class requirement. The remaining *Ss* were employees of the California State Department of Public Health who volunteered to serve as judges. The judgments were made on an 11-point scale ranging from 1 (highly unfavorable to the water supply) to 11 (highly favorable to the water supply). On the basis of mean judgments and Q values, 25 statements were selected to comprise the final scale. These statements with their scale and Q values are presented in Table 1.

The second scale constructed was specific to the taste of water. The 95 statements judged in construction of the attitude taste scale (ATS) were drawn

TABLE 2

SCALE VALUES AND Q VALUES FOR ITEMS SELECTED FOR THE ATTITUDE TASTE SCALE (ATS)

Scale value	Q	Item
1.13	.62	I would never drink our water under any circumstances.
1.24	.96	Our drinking water tastes so bad that I can't stand to keep it in my mouth.
1.65	1.64	I have never tasted such foul water in my life.
1.95	1.17	The taste of our drinking water is so bad that I could only drink it in an emergency
2.16	1.26	The taste of our water is bad in just about every way.
2.47	1.61	Bad tasting water, such as the kind we are now getting, should not be tolerated.
2.78	1.56	I certainly hate to drink our water.
3.36	1.88	The drinking water around here has a disagreeable taste.
4.33	1.30	Our water seems to taste a little bad.
4.81	.96	There is something about our water that gives it a mildly unpleasant taste.
5.00	.69	I am slightly disappointed in the taste of the water which we are now getting.
5.29	1.72	The taste of our water is not bad; it could be better.
5.87	.81	I can't decide whether I like or dislike the taste of our water.
6.00	.65	The taste of our drinking water is neither good nor bad.
6.42	1.55	I have no problems with the taste of our water.
6.62	2.20	Our water will not harm anybody.
6 89	1.58	The taste of our drinking water is pretty fair.
7.67	1.37	The taste of our water seems good.
8.04	1.67	Our water tastes good. I am happy to have it for my drinking water.
8.77	1.57	I drink our water a lot and enjoy it.
8.89	1.60	I find our drinking water to be very satisfying.
9.18	1.58	We're lucky that the taste of the water we get is very good.
9.82	1.41	Our water tastes especially good.
10.34	1.22	The taste of our water is excellent.
10.57	1.20	The taste of our water is perfect in every way.

from the same sources as those for the GAS, and the same procedures and Ss were used in constructing the scale. The items selected for the ATS are shown in Table 2.

A third scale, again specific to the taste of water, was constructed from 105 adjectives taken from the above-mentioned sources and scaled by the same judges and procedures. The adjectives selected for the adjective attitude taste scale (AATS), along with their scale and Q values, are shown in Table 3.

The Interview

An interview was constructed which contained several questions about length of residence and previous water experience, and also contained two unstructured questions dealing with the respondent's evaluation of the community water. One was a question about the water in general, and the other was similar in form but specific to taste. Responses to these questions were recorded verbatim and then coded into favorable, neutral, and unfavorable categories for purposes of comparing attitude scale scores with free responses about the community water.

Studies Using the Scales

Six California towns were selected as sites for determining the performance characteristics of the scales and their relationship to one major factor predicted to be a determinant of potability—total mineral content. The water supplies for these towns contained 50, 82, 388, 1,140, 1,194, and 1,401 milligrams per liter of TDS. The TDS levels were constant throughout each community over the year. No water supply studied had a record of odor problems, and none contained dissolved constituents, other than common minerals, judged capable of producing noticeable taste.

Sampling procedures and circumstances of interview and scale administration varied somewhat from one community to another, but were constant within communities. Respondents in Town B were women selected on the basis of house numbers in a middle-to-low-middle economic area. Respondents in Town F, again all women, were selected by modified quota sampling on the basis of house numbers. The interview and then the attitude scales were administered in the respondents' homes. Respondents in both towns were told that the survey was a study of "people's opinions about their community water supply."

More refined sampling procedures were used in Communities A, C, D, and E. A detailed account of the sampling and survey methods has been given elsewhere (Bruvold, 1965). In each community a sample of residences was randomly selected from the area under study. A census was then made of the selected residences to form a pool of respondents for each town. Respondents were subsequently selected at random from each pool. Eighty-eight percent of the respondents selected from Town D and 95% of those selected from Towns A, C, and E came to the test site (a local school) and participated in the study. Respondents were given the GAS, ATS, and AATS, in that order, as part of a battery consisting also of the interview schedule and a psychophysical procedure involving water samples. The attitude scales followed the interview and preceded the psychophysical task. The relationship between attitude scale scores and psychophysical responses will be the topic of a subsequent report.

TABLE 3

SCALE VALUES AND Q VALUES FOR ITEMS SELECTED FOR THE ADJECTIVE ATTITUDE TASTE SCALE (AATS)

Scale value	Q	Adjective
1.24	.92	horrible
1.36	1.23	sickening
1.43	1.32	terrible
1.68	1.09	disgusting
1.78	2.15	unfit
1.94	1.33	awful
2.09	1.48	good for nothing
2.11	1.76	bad
2.46	1.68	undesirable
2.69	1.94	disagreeable
2.71	1.90	unsuitable
2.87	1.65	unpleasant
3.14	1.29	poor
3.54	1.51	inferior
3.72	1.53	dislikable
4.25	1.51	below par
4.92	1.29	can be tolerated
5.95	1.18	passable
5.97	1.52	not so bad
5.98	.88	fair
6.09	.60	average
6.10	.72	ordinary
6.26	1.04	O.K.
7.94	1.38	likable
8.04	2.10	fine
8.23	1.82	preferable
8.29	1.75	appealing
8.63	1.82	tasty
8.78	1.45	good
9.19	2.08	desirable
9.45	1.74	enjoyable
9.96	1.94	great
10.00	1.99	delightful
10.57	1.60	delicious

RESULTS

Reliabilities of the Scales

To assess reliability of the scales specific to taste, the correlation between ATS and AATS was determined assuming that these were equivalent forms of the same scale. The first line of Table 4 presents these correlations separately by community and for all six com-

TABLE 4

CORRELATIONS AMONG THE ATTITUDE SCALES AND UNSTRUCTURED INTERVIEW QUESTIONS

	Community							
	A (N = 49)	B[b] (N = 19)	C (N = 48)	D[b] (N = 38)	E (N = 51)	F[b] (N = 21)	Combined r	Combined N
ATS–AATS	.93	.83[a]	.95	.92	.94	.94	.94	225
ATS–GAS	.91	.86[a]	.85	.87	.88	.77	.88	225
AATS–GAS	.91	.86	.87	.84	.86	.79	.89	226
ATS–GOE	.59	.69[a]	.84[a]	.62	.64	.51	.72	224
AATS–GOE	.64	.69	.83[a]	.68	.63	.38	.73	225
GAS–GOE	.72	.72	.82[a]	.72	.73	.60	.79	225
ATS–TOE	.58[a]		.74[a]		.77[a]		.75	145
AATS–TOE	.63[a]		.77[a]		.72[a]		.75	145
GAS–TOE	.66[a]		.73[a]		.77[a]		.75	145
GOE–TOE	.67[a]		.71[a]		.68[a]		.71	145

Note.—Abbreviations: ATS = attitude taste scale; AATS = adjective attitude taste scale; GAS = general attitude scale; GOE = open-ended question about evaluation of water in general; TOE = open-ended question about evaluation of water specific to taste.

[a] Based on $N-1$ cases due to one incomplete scale.

[b] The open-ended question specific to taste was not asked in these communities.

munities combined. For five of the six communities the correlations were .92–.95. The correlation for Community B, .83, was below the others but still substantial.

Given the high reliabilities of the ATS and the AATS, the correlation between either of them and the GAS provides a means of estimating the lower boundary of the reliability of the GAS. The reasoning for this is as follows: assuming for the purposes of the limiting case that the true correlation between the GAS and the ATS is 1.00, the formula for correction of attenuation can be solved for the reliability of the GAS.

That is,

$$r_{G,G} = \frac{r^2_{G,T}}{r_{T,T}}$$

TABLE 5

REGRESSION ANALYSIS BETWEEN ATTITUDE TASTE SCALE (ATS) AND TOTAL DISSOLVED SOLIDS (TDS)

Town	A	B	C	D	E	F
TDS	50	82	388	1,140	1,194	1,401
N	49	18	43	38	51	21
M[a]	8.02	7.82	7.59	6.16	6.90	5.88
SD	0.96	1.21	1.19	2.06	1.77	1.89

Source of variation	SS	df	MS	F
Linear regression	109.53	1	109.53	44.89*
Deviation of means from line	19.01	4	4.75	1.98
Between-array means	128.55	5	25.71	10.71*
Within-array means	513.85	214	2.40	
Residual from line	532.85	218	2.44	
Total	642.38	219		

[a] Higher means indicate a more favorable evaluation of the water.

* $p < .001$.

TABLE 6

REGRESSION ANALYSIS BETWEEN ADJECTIVE ATTITUDE TASTE SCALE (AATS) AND TOTAL DISSOLVED SOLIDS (TDS)

Town	A	B	C	D	E	F
TDS	50	82	388	1,140	1,194	1,401
N	49	19	43	38	51	21
M[a]	8.01	7.96	7.56	5.86	6.73	5.41
SD	1.24	1.41	1.20	2.25	1.82	1.83

Source of variation	SS	df	MS	F
Linear regression	160.77	1	160.77	57.01*
Deviation of means from line	26.34	4	6.58	2.39
Between-array means	187.20	5	37.44	13.61*
Within-array means	592.09	215	2.75	
Residual from line	618.44	219	2.82	
Total	779.29	220		

[a] Higher means indicate a more favorable evaluation of the water.
* $p < .001$.

where: $r_{G,G}$ = estimated reliability coefficient for the GAS; $r_{G,T}$ = the obtained correlation between the GAS and the ATS; and $r_{T,T}$ = the reliability of the ATS, taken to be the correlation between ATS and AATS.

Application of this formula indicates a lower bound for the reliability of the GAS to be .82. In fact this is a conservative estimate of the lower bound of the reliability coefficient because the true relationship between the GAS and ATS would probably not be 1.00, since the GAS measures factors other than the evaluation of the taste of the water. To the extent that this is true, the reliability of the GAS will be higher than .82 because of the reciprocal relationship between the correlation between the true scores and the reliability of either scale.

Relationships between the Scales and Free Responses

The correlations between the scale scores and free responses to unstructured questions can be seen in rows four through nine in Table 4. The relationships for all communities combined were all in the .70s even

TABLE 7

REGRESSION ANALYSIS BETWEEN GENERAL ATTITUDE SCALE (GAS) AND TOTAL DISSOLVED SOLIDS (TDS)

Town	A	B	C	D	E	F
TDS	50	82	388	1,140	1,194	1,401
N	49	19	43	38	51	21
M[a]	8.06	7.96	7.49	5.65	6.61	5.26
SD	1.18	1.24	1.26	1.99	1.78	1.99

Source of variation	SS	df	MS	F
Linear regression	191.82	1	191.82	72.66**
Deviation of means from line	31.12	4	7.78	3.05*
Between-array means	222.96	5	44.59	17.49**
Within-array means	547.39	215	2.55	
Residual from line	578.53	219	2.61	
Total	770.35	220		

[a] Higher means indicate a more favorable evaluation of the water.
* $.05 > p > .01$.
** $p < .001$.

though only three coding categories were used for the open-ended questions. These data show that the scales had considerable variance in common with evaluation of the water in the respondents' own words.

Validity of the Scales: Potability and TDS

Regression analyses based on the data from all six towns were performed to determine the relationships between the attitude scales and TDS. These analyses are presented in Tables 5, 6, and 7. In all cases the differences among means for the different communities were significant beyond the .001 level. Additional tests show the linearity of relationship to be highly significant in all cases ($p < .001$). In only the case of the GAS was there significant ($p < .05$) departure from the linear-regression line.

Correlation coefficients (Pearson r) between TDS and ATS, AATS, and GAS were $-.41$, $-.45$, and $-.50$, respectively; the corresponding etas were .45, .49, and .54.

DISCUSSION

The high intercorrelations among the scales indicate that a very satisfactory level of reliability was achieved. It is well to note that respondents from the six towns were, because of the sampling methods employed, extremely heterogeneous in age, education, and other characteristics. Thus, a rigorous test of the performance of the scales was provided in terms of understandability, ambiguity of items, and overall consistency of scores.

The substantial correlations found between coded response to unstructured-interview questions and the attitude-scale scores indicate that these very different methods of measuring attitudes toward domestic water yielded similar results. These data, like the reliability data, showed that the attitude scales performed in a satisfactory manner.

The high level of statistical significance obtained in the relationship between the attitude taste scales and TDS confirmed the hypothesis of the study: progressively higher levels of TDS were associated with progressively less favorable attitudes. This result supports the theoretical notion that potability is inversely related to total mineral content in domestic water, and it also supports the validity of the scales constructed to measure consumer attitudes toward the taste of domestic water.

It should be noted here that social-psychological theory and research indicate that individual attitudes may be determined by a number of factors, and that they are usually the result of considerably more than the physical or "objective" characteristics of the attitude object. The attitudes of individual consumers toward the taste of the community water, then, should be determined by the physical characteristics of the water and other factors as well. A theory has recently been developed (Dillehay & Siegel, 1966) which outlines in some detail possible determinants of consumer attitudes toward taste in domestic water. The theory involves a functional view of attitudes that owes much to previous work by Katz and others (e.g., Katz, 1960; Katz & Stotland, 1959; Kelman, 1961; Smith, Bruner, & White, 1956). In broad strokes, two general classes of factors are seen to affect attitudes toward the community water: psychophysical and social psychological. Psychophysical influences include the experience of the water as a function of its physical characteristics (e.g., total dissolved minerals, gasses given off by algae) and the state of sensory mechanisms of the individual (e.g., due to habituation, previous taste experience, physiological sensitivity). The social-psychological class of determinants includes individual taste preferences, perceived social norms referring to the water, the psychological context of the community water supply (e.g., the attractiveness to the individual of other aspects of the community), and the perceived nature of alternatives to the present supply, including what might be done to improve the present water.

The magnitude of the correlations between TDS and attitude scale scores found in the present research, while statistically significant and potentially useful for predicting mean attitude scores for communities, indicates that a considerable portion of the attitude-scale score variance was unaccounted for by TDS. Since mineral content was the only important taste-producing physical constituent in the waters studied, the magnitude of the obtained correlations suggests that sensory and social-

psychological factors just discussed produce considerable variation in attitude-scale scores. Those sensory and social-psychological variables that are amenable to assessment should be systematically measured in future surveys dealing with the relation between mineral content and potability in order to provide a test of the theory just proposed. The measurement of these variables will also aid in recommending an upper limit for TDS in domestic water in order to ensure potability for daily drinking.

An upper limit for total mineral content in domestic water could be recommended with reasonable confidence if, in future surveys conducted as in Towns A, C, and E and covering a sizable number of community water systems, the relationship between TDS and attitude-taste-scale scores turns out to be linear, or nearly so, and sensory and social-psychological determinants of attitude toward mineral taste in water are found to be uncorrelated with TDS. Given linearity and lack of correlation between TDS and additional determinants of attitude, an upper limit for TDS in domestic water could be determined from the univariate regression equation relating mean attitude scores and TDS. The upper limit for TDS would be set by the mean attitude score defining the reasonable limit of potable domestic water.

If the relationship between TDS and mean attitude-taste-scale scores departs considerably from linearity, upper limits for minerals in domestic water might be recommended using a multivariate linear-regression equation in which separate ionic constituents serve as predictor variables. If mean attitude scores for communities did not deviate significantly from multivariate linearity, and sensory and social-psychological determinants of attitudes did not correlate significantly with any of the several predictor variables, upper limits for the several ionic concentrations would also be established by the mean attitude score defining the reasonable limit of potable domestic water. If, however, considerable correlation occurs between sensory or social-psychological determinants of attitude and one or more of the ionic constituents, or with TDS in the simpler situation, the relation between amount of mineral content and measured attitudes would have to be accounted for, separate from these other factors, before upper limits for minerals in domestic water could be accurately recommended.

REFERENCES

BRUVOLD, W. H. Consumer attitudes toward mineral taste in domestic water. Technical Report submitted to Public Health Service (Contract No. PH 86-64-77), 1965.

DILLEHAY, R. C., & SIEGEL, J. P. Attitudes and community water supplies. Technical Report submitted to the Division of Research, California State Department of Public Health, 1966.

EDWARDS, A. Techniques of attitude scale construction. New York: Appleton-Century-Crofts, 1957.

KATZ, D. A functional approach to the study of attitudes. *Public Opinion Quarterly*, 1960, 24, 163–204.

KATZ, D., & STOTLAND, E. A preliminary statement to a theory of attitude structure and change. In S. Koch (Ed.), *Psychology: A study of a science*. Vol. 3. New York: McGraw-Hill, 1959. Pp. 423–475.

KELMAN, H. Processes of opinion change. *Public Opinion Quarterly*, 1961, 25, 57–78.

ONGERTH, H., BRUVOLD, W. H., & KNUTSON, A. The taste of water. *Public Health Reports*, 1964, 79, 351–354.

SMITH, M. B., BRUNER, J. S., & WHITE, R. W. *Opinions and personality*. New York: Wiley, 1956.

(Received June 6, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 96–100

DRIVING RECORD OF NEUROPSYCHIATRIC PATIENTS[1]

MATTHEW W. BUTTIGLIERI
Veterans Administration Hospital, Sepulveda, California

AND

MARIE GUENETTE
Institute of Transportation and Traffic Engineering, University of California, Los Angeles

The driving records were obtained from the California Department of Motor Vehicles for all patients admitted to the neuropsychiatric wards of the Sepulveda (California) Veterans Administration Hospital with active driver's licenses in their possession. The 165-patient sample had accident and violation records which did not differ significantly from the California male driving population. 80% of the sample had no accidents and more than ⅔ of the sample had no more than 1 driving violation during the 3 yr. preceding hospitalization. There was no clear-cut difference in negligent-operator point count between the patient group and the California male driving population except at the high end of the point-count distribution for a very small proportion of patients.

In this highly mechanized age the lack of transportation can be a serious economic handicap to the working individual. In population centers where public transportation is less than adequate, the individual is forced to depend upon the automobile for crucial aspects of his livelihood. In few sectors of our society is the lack of transportation more vital, socially and economically, than in the rehabilitation of the recovered mental patient.

Understandably, most authorities are reluctant to deprive the individual of the driving privilege. At the same time, the identification and reduction of high-risk automobile drivers is a growing concern especially to those responsible for highway safety. Most licensing agencies rely exclusively on occasional dramatic highlighting of deficient driving skills. Some require the applicant for the driver's license to answer questions related to presumed high-risk driving conditions, for example, mental illness, epilepsy, dizzy spells, paralysis, etc. Few, if any, have a systematic procedure to check upon the veracity of the answers to determine with certainty the presence or current status of the presumptive high-risk condition.

The literature is full of studies of the human element in driving, and many of the psychological characteristics found to be related to accident susceptibility are those which one would expect to find accentuated in the psychiatric patient. However, curiously enough, there has been no study of driving characteristics with special reference to the mental hospital patient. Despite this lack of hard data, the individual with a history of mental illness is uniquely singled out for special scrutiny where the driving privilege is concerned.

It is common hospital practice to confiscate the driver's license along with other personal belongings of the patient at the time of admission. Lamb (1959) has cited cases to illustrate that latent tendencies to act out with the automobile may be activated by conflict situations generated during the course of psychotherapy, and he recommends that the therapist should take action to forestall the occurrences of accidents in therapy. The danger of accidents becomes an important consideration in the management of the hospitalized mental patient when the question of passes, night hospitalization, trial visit, etc., comes up for decision by treatment teams. All too often, however, the actions and the decisions are made indiscriminately because of the lack of relevant data.

The patient who is unable or unwilling to

---

[1] This study is part of a 3-year investigation conducted by the Institute of Transportation and Traffic Engineering in conjunction with the Sepulveda Veterans Administration Hospital supported by a research grant from the Accident Prevention Division of Bureau of State Services, Department of Health, Education and Welfare.

The authors gratefully acknowledge the cooperation of the Western Research Support Center who helped with the data processing and statistical analysis.

TABLE 1

CHARACTERISTICS OF PATIENT SAMPLE

Characteristic	Frequency	Percentage
Age		
Under 25	6	3.63
25–34	36	21.81
35–44	64	38.78
45–54	43	26.06
55–64	7	4.24
65–74	8	4.84
75–84	0	.0
85–94	1	.6
Mean age = 42.09		
Median age = 40.75		
Marital status		
Single	43	26.06
Married	74	44.85
Separated	4	2.42
Widowed	6	3.64
Divorced	38	23.03
Sex and race		
Male white	139	84.24
Male Negro	5	3.03
Male, not stated	11	6.67
Female white	10	6.06
Diagnosis		
Schizophrenic	90	54.55
Psychotic disorders (other)	10	6.06
Chronic brain syndrome	12	7.27
Psychoneurotic disorders	35	21.21
Personality disturbance	9	5.45
Other	9	5.45

await the favorable decision with respect to the return of his driver's license by hospital authorities has several alternatives. He can, of course, take the chance of driving without a license despite the potentially serious consequences. The recidivist patient learns to insure for himself the convenience of the driving privilege by leaving his license at home or with a relative thus avoiding its confiscation upon hospital admission. Finally, the knowledgeable patient can, at the earliest opportunity, report his license as lost and thus obtain a replacement at modest cost to himself. For those instances where there is a serious question of the patient's fitness to drive, some tightening up of these loopholes seems in order. However, since the withholding of the driving privilege can have most serious economic repercussions on the patient and his family, recommendations in this area should be made only after the most careful study and the establishing of a practical set of criteria.

This study, which is part of a 3-year investigation of the effect of drugs on skills related to driving, examines the driving records of the neuropsychiatric patient and how these relate to comparison groups.

METHOD

The subjects (*Ss*) of this study consisted of patients admitted to neuropsychiatric wards at the Sepulveda (California) Veterans Administration Hospital from August 1964 through July 1965. All those patients with active California driver's licenses in their possession at the time of admission were considered for inclusion in this study. Identifying information from the licenses was sent to the California Department of Motor Vehicles on the Driver's License Information Request Form. These forms were subsequently returned with complete information with respect to the record of accidents and violations in the 3-year period immediately preceding hospital admission.

RESULTS

A total sample of 165 different patients was collected. This represents approximately 29% of the total psychiatric admissions for the 12-month period of the study. Table 1 summarizes characteristics of the patient sample in terms of age, marital status, sex, race, and diagnosis. The *Ss* varied in age from 23 to 85 with a mean age of 42.1. The

TABLE 2

DISTRIBUTION OF ACCIDENT FREQUENCIES FOR PATIENT SAMPLE AND RANDOM SAMPLE OF CALIFORNIA DRIVERS

Accidents	California drivers' percentages	Patient frequencies	Expected patient frequencies[a]
None	78.66	133	129.79
1	17.49	31	28.85
2	3.20	1	5.28
3	0.53	0	.87
4	0.10	0	.17
5	0.01	0	.02
6 or more	0.01	0	.02
Total	100%	165	165.00

Note.—$\chi^2 = 4.75$; $.10 > p > .05$.
[a] Since small expected frequencies may lead to a violation of the fundamental assumptions underlying the use of χ^2, the categories of two accidents or more were combined (Maxwell, 1964).

modal patient is a 40-year-old married male, white schizophrenic.

Table 2 presents the accident distributions for the patient sample and for a random sample of the California male driving population (Department of Motor Vehicles, State of California, 1965). The male population ($N = 86{,}717$; mean age 40.3, median age 38.8) was used as the most appropriate comparison group, since 94% (155 patients) of the patient sample was male. Eighty percent (133 patients) of the sample had no accidents in the 3 years prior to admission. The mean accident rate for the patient sample is .20; the mean accident rate for male California driving sample is .26. The next step was determining whether this record of accidents is typical of the California driving population. The chi-square test of goodness of fit was utilized to test whether the discrepancies between the observed (patient) frequencies and the expected frequencies (derived from the California driving population) could reasonably be attributed to chance fluctuations in the observed data (Maxwell, 1964). Chi-square analysis yielded a value of 4.75 which, with 2 degrees of freedom, is not significant at the .05 level.

Table 3 presents the distributions of violations for the patient sample and for the California male driving population. Almost one-half (76 patients) had no violations and over two-thirds (115 patients) had no more than one driving violation during the 3 years preceding psychiatric hospital admission. The mean violation rate for the patient sample is 1.34; the mean violation rate for the male California driver is 1.10. The chi-square value of 5.15 with 5 degrees of freedom is not significant at the .05 level. Therefore, it seems safe to conclude that with respect to accidents and violations no real difference is found between the observed frequencies in the patient sample and frequencies to be expected were a sample drawn randomly from the California male driving population.

All recorded violations were next examined in order to determine the "negligent-operator point count" for each patient in the sample. This is a designation used by the California Department of Motor Vehicles and is computed in the following manner: Each recorded accident results in 1 point and each recorded conviction (driving violation) results in 1 point with the exception that a recorded conviction results in 2 points if it involves one

TABLE 3

DISTRIBUTION OF VIOLATION FREQUENCIES FOR PATIENT SAMPLE AND RANDOM SAMPLE OF CALIFORNIA DRIVERS

Violations	California drivers' percentages	Patient frequencies	Expected patient frequencies[a]
None	49.87	76	82.28
1	24.44	39	40.33
2	11.84	24	19.54
3	5.94	9	9.80
4	3.13	4	5.16
5	1.82	5	3.00
6	1.15	2	1.89
7	.68	2	1.12
8	.39	0	.64
9 or more	.74	4	1.12
Total	100%	165	165.1

Note.—$\chi^2 = 5.15$; $.50 > p > .30$.
[a] In this table the number of violations of five and above were combined in order not to violate the assumptions underlying the use of χ^2.

TABLE 4

DISTRIBUTION OF NEGLIGENT-OPERATOR POINT COUNTS FOR PATIENT SAMPLE AND RANDOM SAMPLE OF CALIFORNIA MALE DRIVERS

Negligent-operator point count	California male drivers' point count (percentages)	Patient frequencies	Expected patient frequencies[a]
0	46.18	67	76.2
1	24.58	38	40.6
2	13.21	21	21.8
3	7.16	13	11.8
4	3.76	9	6.2
5	2.14	5	3.5
6	1.34	3	2.2
7	0.69	4	1.1
8	0.42	1	0.7
9	0.22	1	0.4
10 or more	0.30	3	0.5
Total	100%	165	165.0

Note.—$\chi^2 = 11.67$; $.05 > p > .02$.
[a] In this table the negligent-operator point counts of 5 and above were combined in order not to violate the assumption underlying the use of χ^2.

of the six violations (e.g., misdemeanor drunk driving, hit and run, reckless driving, etc.) designated in the California Vehicle Code. A driver who obtains at least 4 points during any 12-month period, or 6 points during any 24-month period, or 8 points during any 36-month period is defined as a "negligent operator." These drivers are subsequently called in for a hearing during which details of his accidents and violations are examined and a decision is reached with respect to revocation or suspension of the driving privilege.

Table 4 presents the frequency distribution by negligent-operator point count for the patient sample and the male ($N = 86,726$) sample of the California driving population. The chi-square value of 11.67 with 5 degrees of freedom is significant at beyond the .05 level of confidence. Table 4 indicates that five patients have a point count of 8 or more, three of these have a point count of 10 or more. These would definitely be considered negligent operators and in the theoretically perfect system, each would be immediately summoned to a hearing for the purpose of deciding on possible suspension or revocation of the driving privilege.

Table 5 indicates that when these five cases are removed from the distribution along with all those in the male California driving population with a negligent-operator point count of 8 or more, the resulting χ^2 is reduced to the point where it is statistically not significant. From the data there appears to be no clear-cut difference between the patient group and the California male driver group, except at the high end of the point-count distribution for a very small proportion of patients.

TABLE 5

DISTRIBUTION OF NEGLIGENT-OPERATOR POINT COUNTS FOR PATIENT SAMPLE AND RANDOM SAMPLE OF CALIFORNIA MALE DRIVERS WITH "NEGLIGENT OPERATOR" CASES REMOVED

Negligent-operator point count	California male drivers' point count (percentages)	Patient frequencies	Expected patient frequencies[a]
0	46.61	67	74.66
1	24.81	38	39.70
2	13.33	21	21.33
3	7.23	13	11.57
4	3.80	9	6.08
5 or more	4.17	12	6.67

Note.—$\chi^2 = 6.70$; $.30 > p > .20$.
[a] In this table the negligent-operator point counts of 5 and above were combined in order not to violate the assumptions underlying the use of χ^2.

CONCLUSION AND DISCUSSION

The results of this study indicate that most of the patients admitted to the neuropsychiatric wards of the Sepulveda (California) Veterans Administration Hospital with active driver's licenses in their possession have accident and violation records which are found not to differ significantly from the California male driving population. This finding is consistent with the studies that indicate that when Veterans Administration patients are returned to the community, they commit fewer acts of violence and fewer felonies than average males who have never been in a mental hospital. Some additional degree of support is indicated from a study by Davis and Coiley (1959). They conducted a psychological and medical study of a group of motor-vehicle drivers whose accident records suggested that they were accident prone and compared these with a group of drivers with records good enough to compete in safe-driving awards. They state that the most striking finding in their study is that all of the eight *Ss* in their sample ($N = 126$) who had suffered from a definite nervous or mental illness were in the "safe" rather than in the "accident-prone" group.

The results also suggest that any further restriction of the mental patient's driving privilege will not, in many instances, be justified by his driving record. For various reasons many individuals with a history of mental illness do not drive. In many cases this is a result of the lack of assertiveness in the individual whose license has lapsed during his hospitalization. Undoubtedly, among these are individuals who could be rehabilitated more effectively if they obtained and exercised their driving privilege. This could be accomplished without adding to the already significant risks on the highway.

On the other hand, 3% of the sample studied had driver records which were serious enough to classify them as negligent operators. These people warrant further study in order to more precisely identify the psychologically definable characteristics of the mental patient which may be expected to cause accidents. There is an apparent reluctance on the part of the judiciary to exclude from licensure an individual who is ill or disabled, unless it is demonstrated and supported by factual data that his impairment under the normal conditions of use affects his driving ability. Despite this, it is current hospital practice to presume that need for psychiatric hospitalization is sufficient cause to restrict the driving privilege. With the increasing usage of the night and day hospital plans and the increasing practice of reducing the cultural discontinuities imposed by traditional hospitalization, there is a need to square the practice of restricting driving with the social and economic necessity to drive. The most economical manner of accomplishing this would involve closer cooperation between hospital authorities and the driver-improvement analysts in state motor-vehicle agencies for the purpose of early identification and prompt selective restriction of the mental patients who are serious highway risks.

REFERENCES

BROWN, P. L., & BERDIE, R. F. Driver behavior and scores on the MMPI. *Journal of Applied Psychology*, 1960, **44**, 18–21.

DAVIS, R. D., & COILEY, P. A. Accident-proneness in motor vehicle drivers. *Ergonomics*, 1959, **2**, 239–246.

DEPARTMENT OF MOTOR VEHICLES, STATE OF CALIFORNIA. *The 1964 California driver record study*, 1965, Part 2.

LAMB, H. R. Automobile accidents associated with psychotherapy, *American Journal of Psychotherapy*, 1959, **13**, 899–905.

MAXWELL, A. E. *Analysing qualitative data*. London: Methuen & Company, Ltd., 1964. Pp. 97–99.

(Early publication received May 16, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 101-106

INTERRATER RELIABILITY IN SITUATIONAL TESTS

JOHN M. GREENWOOD AND WALTER J. McNAMARA

International Business Machines Corporation, Armonk, New York

This study was conducted to determine the degree of interrater reliability in situational tests and to determine the relative effectiveness of professional and nonprofessional evaluators in this type of situation. The results indicate that the reliability of observer ratings and rankings are reasonably high in several different situational tests. Of particular significance is the finding that adequate reliability can be obtained from the use of nonprofessional evaluators in business-oriented situational tests.

In today's society, composed of large groups of individuals, a hierarchical structure of organization is prevalent in business, governmental and religious organizations, etc. The selection of individuals to move up the ladder to positions of greater responsibility, particularly in business organizations, has traditionally relied on the recommendations of the man's immediate superior. This approach continues to be widely used; but in recent years, in an attempt to improve on the traditional approach, the use of objective tests in managerial selection has been studied. Research studies utilizing objective tests have been conducted by several organizations (Laurent, 1961, 1962; Mahoney, Jerdee, & Nash, 1961; Spitzer & McNamara, 1964). More recently, a few business organizations have undertaken programs utilizing situational tests in addition to written objective tests (Bray, 1962; Educational Testing Service, 1964). Situational tests have been succinctly described (Flanagan, 1954) as follows:

> Situations are selected to be typical of those in which the individual's performance is to be predicted. . . . (Each) situation is made sufficiently complex that it is very difficult for the persons tested to know which of their reactions are being scored and for what variables. There seems to be much informal evidence (face validity) that the person tested behaves spontaneously and naturally in these situations. . . . It is hoped that the naturalness of the situations result in more valid and typical responses than are obtained from other approaches [p. 462].

Therefore, although situational tests do not elicit the individual's specific "on-the-job" behavior, they do place the individual in a situation whereby some basic managerial characteristics will have to be displayed in order to carry out the situational task. There are two types of situational tests commonly used: group exercises—the participants are placed in a situation in which the successful completion of the task necessitates interaction among the participants; and individual exercises—the participant completes the task independently, as in the In-Basket (Lopez, 1966; Educational Testing Service, 1964).

In the group situational test each participant is usually evaluated by two or more observers. Accordingly, a practical problem encountered in the use of this technique is the degree of rater agreement. Without appropriate reliability of observer evaluations, the potential validity and usefulness of situational tests in managerial selection or other similar situations is of course questionable.

Although primarily focusing upon the testing and formulation of multiple methods in personality assessment, Taft (1959) discusses some issues concerned with the "background" of assessors who have been used in various assessment programs. He reports that studies tend to indicate that the highest validations are achieved by those judges who have the greatest familiarity with the types of people who are being assessed. He goes on to imply that, all else being equal, the best assessors for predicting existing criteria are those who are partially contaminated with the experiences, standards, and outlooks of those who are evaluated in the assessment program.

This would indicate that line managers or executives should be used as evaluators in managerial assessment programs. On the other hand, the problems involving judging biases are numerous in situational assess-

ment programs. Perhaps assessors who have professional training in assessment could better objectify the participant's performance and obtain greater interrater reliability than so-called nonprofessional managerial officials.

The present study, which is part of a larger project pertaining to managerial selection, analyzed the degree of interrater reliability of situational tests and presents some results pertaining to the use of professional and nonprofessional assessors.

PROCEDURE

As part of a 2-day personnel assessment program, individuals were assessed by situational (group and individual) tests and by traditional paper-and-pencil tests. In each 2-day program, 12 participants were evaluated by a team of 4 observers. In each situational test, however, 6 of the participants were evaluated by only 3 observers. Thus, in one exercise, the 6 participants would be evaluated by Observers A, C, and D; in another exercise by A, B, and C; and in a third by B, C, and D. At the completion of an exercise each observer independently rated and ranked each participant in the group.

Variables

Observer ratings and rankings. First, each observer rated the 6 individuals on a 5-point scale: (1) much more effective than most in the group, (2) somewhat more effective, (3) about as effective as most in the group, (4) relatively ineffective, (5) ineffective. Effectiveness was based on specific characteristics (i.e., aggressiveness, oral communications skill, etc.). Second, each observer ranked the 6 group members from 1 (best) to 6 (poorest) on their overall effectiveness in the group.

Group situational exercises. In this study three different types of group exercises were used. These three situational tests were all related in some manner to the functioning of a business concern, although in each exercise there were distinct managerial abilities and various areas of knowledge being measured.

1. Leaderless Group Discussion: Each participant was required to make a 5-minute oral presentation of a candidate for promotion, and then subsequently each participant defended his candidate in a group discussion with the other participants. The roles given the participants were well defined and structured.

2. Task-Force Committee: Participants were given data from which they were to form a group recommendation to the president of "X" company on which of three alternate courses of action to take for expansion of the company's activities. This exercise was similar to the leaderless except for less-structured roles.

3. Manufacturing: The participants were required to work together as a group and operate a manufacturing company in an effective manner. They were required to purchase materials, manufacture the product, and sell it back to the market. Included in this exercise was a product forecast, fluctuation of prices for raw materials, and completed products, etc. There were no assigned roles for the participants, and both verbal and physical activity were prevalent in this exercise.

Subjects. During the past 2 years, several divisions of a large electronics company have undertaken a personnel assessment program. For statistical purposes, only those divisions which have evaluated at least 36 individuals are reported. Four divisions (A, B, C, D) of the company have conducted six different programs. For Division A, there were three different assessment programs separated by a considerable period of time, and, there-

TABLE 1

RELIABILITY COEFFICIENTS OF OBSERVER RATINGS AND RANKINGS

Division	Task Force		Leaderless		Manufacturing	
	Ratings Mean r (Z)[a]	Rankings Mean r (Z)[b]	Ratings Mean r (Z)	Rankings Mean r (Z)	Ratings Mean r (Z)	Rankings Mean r (Z)
A (N = 144)						
I (N = 48)	.75	.81	.70	.74	.88	.84
II (N = 48)	.62	.70	.57	.62	.77	.77
III (N = 48)	.86	.82	.82	.83	.88	.89
B (N = 48)	.74	.75	.58	.60	.77	.83
C (N = 60)	.65	.65	.66	.68	.74	.81
D (N = 36)	.84	.78	.48	.48	.49	.49

[a] Pearson product-moment correlations were used to determine the degree of rater agreement for an experimental group (in each experimental group three observers independently evaluated six participants who were completing a particular situational test). In order to determine an estimate of rater agreement for an entire division program, the interrater reliabilities obtained from experimental groups in a particular division were then averaged (after converting to Fisher Z transformations).

[b] The same statistical procedure was used to determine reliability of ranking except that in each group the observers ranked the participants and Kendall's coefficient of concordance was used to obtain the degree of ranking agreement for an experimental group.

TABLE 2

MEANS AND STANDARD DEVIATIONS OF OBSERVER RATINGS

Division	N	M	SD
Task Force			
A	144	3.088	1.001
B	48	2.742	1.054
C	60	2.777	.986
D	36	2.964	1.009
Total	288	2.950	1.020
Manufacturing			
A	144	3.039	1.088
B	48	2.873	1.077
C	60	3.047	.971
D	36	2.883	.896
Total	288	3.007	1.066
Leaderless			
A	144	2.864	.958
B	48	2.658	.883
C	60	2.855	1.090
D	36	2.686	.842
Total	288	2.799	2.966

fore, roman numerals I, II, and III indicate results from these three programs. In five of these programs (AI, AII, AIII, B, and C), the individuals assessed came from the same functional area and were from lower- or middle-management-level positions. In the D divisional program, the individuals being assessed were from various functional areas of the company, and were generally from middle- to upper-level managerial positions.

Evaluators. The individuals who served as evaluators can be classified into two groups: professional (psychologists or sociologists) or nonprofessional (high-level experienced managers, but with only minimum training in assessment techniques). The evaluators in Programs A, B, and C were at least two managerial levels above the participants, but in Program D the nonprofessional evaluators were only slightly above the participants in managerial status.

RESULTS

In Table 1 are presented the mean reliabilities of observer ratings and rankings according to divisional program and situational exercise. In this table, N refers to the number of participants. As indicated above, the participants were evaluated in groups of six by three observers. Thus the mean correlation rating of .75 for the Task Force for Division AI is based on 24 correlations: 8 groups and 3 sets of observers. These results indicate that the reliability of observer ratings and rankings are reasonably high in most cases; the exception to this general statement are the reliabilities for Division D for Leaderless and Manufacturing exercises. The rank-order coefficients may be somewhat spuriously higher than Pearson r coefficients (due to algebraic differences).

A comparison of observer-rating and observer-ranking reliabilities indicates that these two methods are approximately equivalent regardless of situational exercise or divisional program.

In Table 2 are presented the means and standard deviations of the observer ratings of participants according to divisional program and group situational exercise. As indicated above, these ratings were based on a 5-point scale, 1–5. F tests were undertaken (Table 3) to test whether there were significant variances of observer-rating distributions among the divisional programs. For all three situational exercises the results were not significant. Therefore, the distribution of ratings among the divisions was generally consistent with one another.

Due to the limited amount of research data available in this area, it is difficult to evaluate the adequacy of these observer reliabilities to what might be obtained in other similar programs. Some data are available, however, from an AT&T study (Grant,

TABLE 3

ANALYSIS OF VARIANCE AMONG DIVISIONAL PROGRAMS

Division	SS	df	MS
Task Force[a]			
Between	6.62	3	2.21
Within	288.61	284	1.02
Total	295.23	287	
Manufacturing[b]			
Between	1.13	3	.38
Within	314.5	284	1.11
Total	315.63	287	
Leaderless[c]			
Between	3.57	3	1.19
Within	279.1	284	.98
Total	282.77	287	

[a] $F = 2.16$; $F_{.95} = 2.64$.
[b] $F = .34$; $F_{.95} = 2.64$.
[c] $F = 1.22$; $F_{.95} = 2.64$.

TABLE 4

COMPARISON OF OBSERVER RELIABILITY IN TWO SIMILAR PROGRAMS

Division	Present study (N = 288)	AT&T (N = 355)
Task Force		
Overall rating	.70	No data available
Overall ranking	.71	
Leaderless		
Overall rating	.66	.75
Overall ranking	.64	.75
Manufacturing		
Overall rating	.74	.60
Overall ranking	.75	.69

1964), and a comparison of observer agreement is presented in Table 4. Analysis was made to compare these two independent populations as to whether they have equivalent interrater agreement (Walker & Lev, 1953, p. 255). The analysis indicates that the AT&T reliabilities were significantly higher ($p < .05$) for the Leaderless exercise, and the present study reliabilities are significantly higher ($p < .01$) for the Manufacturing exercise. However, in both studies the interrater agreement is of a magnitude to warrant the continued use of this type of situational testing.

In comparing the three types of situational exercises, it will be seen that there is a tendency for the Manufacturing exercise to have higher reliabilities than the two other exercises. The one exception to this general statement is Division D.

After an initial examination of the observer reliabilities, it was conjectured that in those situational exercises for Division D, where low reliability of observer ratings appeared, it might be partially due to the variation in the professional experience of the observers. The basic rationale of this hypothesis was that the use of both professional and nonprofessional evaluators (heterogeneous) in the same situational assessment exercise might consistently lower the reliability of observer agreement compared to a group of homogeneous evaluators.

From an examination of the data, 60 individuals were identified who had been evaluated by two professional and two nonprofessional evaluators. Since 12 participants were evaluated by the same 4 evaluators, the ratings of these groups of 12 participants are presented individually in Tables 5 and 6.

In Table 5, each case (N) represents two evaluators rating a participant. The reason for the inequalities of cases within an exercise or among groups was due to procedural arrangements. For example, in one group exercise, 6 participants might have been rated by 2 nonprofessionals and 1 professional evaluator, resulting in 12 comparisons of nonprofessional with professional and 6 comparisons of nonprofessional with nonprofessional.

In Table 5, the product-moment correlations between nonprofessional and nonprofes-

TABLE 5

COMPARISON OF RATING RELIABILITIES FOR PROFESSIONAL VERSUS NONPROFESSIONAL OBSERVERS

Group	Task Force			Leaderless			Manufacturing		
	NP vs NP	P vs NP	P vs P	NP vs NP	P vs NP	P vs P	NP vs NP	P vs NP	P vs P
D_1 (N = 12)	.21 (N = 6)	.80 (N = 24)	.91 (N = 6)	.71 (N = 6)	.12 (N = 24)	.43 (N = 6)		.20 (N = 24)	.18 (N = 12)
D_2 (N = 12)	.78 (N = 6)	.57 (N = 24)	.68 (N = 6)	−.05 (N = 6)	.47 (N = 24)	.87 (N = 6)		.23 (N = 24)	.55 (N = 12)
D_3 (N = 12)	.89 (N = 24)	.80 (N = 12)		.71 (N = 12)	.63 (N = 24)		.68 (N = 24)	.75 (N = 12)	
B (N = 12)	.86 (N = 6)	.79 (N = 24)	.58 (N = 6)	.78 (N = 6)	.81 (N = 24)	.75 (N = 6)	.45 (N = 6)	.70 (N = 24)	.88 (N = 6)
C (N = 12)	.63 (N = 6)	.75 (N = 24)	.97 (N = 6)	.81 (N = 12)	.78 (N = 24)		.81 (N = 12)	.71 (N = 24)	
Total N = 60									

Note.—Abbreviations: NP = nonprofessional; P = professional.

TABLE 6

PROFESSIONAL VERSUS NONPROFESSIONAL OBSERVER RATINGS

Division	Nonprofessional			Professional			t
	M	SD	N	M	SD	N	
Task Force							
D_1	3.17	1.10	18	3.06	1.05	18	.69
D_2	2.78	0.70	18	2.83	0.94	18	.24
D_3	3.00	1.37	18	2.89	1.09	18	.87
B	3.00	1.48	18	2.83	1.30	18	.36
C	2.72	1.33	18	3.00	1.28	18	1.28
Leaderless							
D_1	2.78	1.06	18	2.61	1.15	18	.57
D_2	2.50	0.79	18	2.72	1.25	18	.76
D_3	3.17	1.20	18	2.67	0.97	18	2.33*
B	2.61	1.14	18	2.72	1.10	18	.61
C	2.83	1.27	24	2.75	1.00	12	.41
Manufacturing							
D_1	3.08	0.91	12	2.92	1.10	24	.47
D_2	2.50	1.17	12	2.67	1.20	24	.45
D_3	3.33	1.07	12	2.96	1.34	24	1.78
B	3.17	1.30	18	2.83	0.98	18	1.36
C	3.13	1.13	24	3.00	1.04	12	.66

* $p < .05$.

sional ratings, professional and nonprofessional ratings, and professional and professional ratings are presented. Appropriate t tests (two-tailed tests) were computed (Table 6) to determine whether there were significant differences between the mean ratings of the two types of observers for the same group of participants (Walker & Lev, 1953, p. 151).

In some instances the interrater reliability is alarmingly low (Table 5), but when one compares in a given situation the heterogeneous group (professional and nonprofessional) reliabilities to the homogeneous group (professional and professional, or nonprofessional and nonprofessional) reliabilities, they generally appear equivalent. Also, neither homogeneous group consistently appears to have greater interrater reliability than the other homogeneous or heterogeneous groups.

The results in Table 6 indicate that neither type of evaluator consistently gave higher or lower ratings to the participants (in Table 6, low numerical ratings indicate high participant effectiveness), and the distribution of rating scores is about equivalent for both types of assessors. These results suggest that there are no significant differences between professional and nonprofessional raters, except one instance where the results did reach significance (pure chance probability).

A possible explanation for the lower observer agreement in the Division D program is (perhaps) due to two distinctive features of this program in comparison to the other programs: first, the participants assessed were from various functional areas; and second, the nonprofessional observers were practically at the same managerial level as the participants. The data obtained in this study cannot adequately dilineate which one or combination of factors might have some bearing upon lower rater agreement in the Divisional D program.

DISCUSSION

From the above results, one should not assume that nonprofessional personnel, who are without substantial business experience or lack any training in evaluating participants, should be employed in assessing personnel. Furthermore, as in this study, the instructions to the evaluators, whether professional or nonprofessional, should be explicit as to the specific type of behavior to be evaluated, examples provided as to kinds of behavior that may be expected to be elicited by the specific situational exercises, and standardized rating forms provided for their use. However, the results tend to indicate that the selection of evaluators for this type of assessment need not be based on whether they have completed an extensive training program or have professional experience in personnel assessment. Perhaps an exception to this rule would be the evaluation of top echelon executives in a company where only independent professional evaluators could objectively evaluate the participants.

The results of this study suggest that reasonable rater reliability can be obtained in situational tests commonly used to assess business and industrial personnel regarding their potential for advancement. These results, of course, in no way indicate the validity of situational tests in predicting managerial potential or success.

The results further suggest that both observer ratings or rankings of participants in situational tests generally have adequate reliability as methods for evaluating performance in situational exercises.

The comparison between professional and nonprofessional evaluators indicated that their

ratings are about equivalent for a specific group of participants being assessed. Therefore, an expectation that ratings by nonprofessional assessors in such exercises would differ from professional assessors—due to differences in experience, training, and awareness of common judging biases—seems unwarranted. In general, it can be concluded that either type of evaluator can be used interchangeably as observers. There might still be judging biases associated with ratings of situational tests, but both types of evaluators should give either equally valid or invalid ratings.

Research should be undertaken to show whether a comparison between evaluators and participants on certain socioeconomic or status indicators would increase or diminish the reliability and validity of observer ratings.

REFERENCES

BRAY, D. W. The management progress study. *The Conference on Executive Study: Identifying management talent.* Princeton, N. J.: Educational Testing Service, 1962. Pp. 47–73.

EDUCATIONAL TESTING SERVICE. *Management games in selection and development, The Executive Study Conference, May 5 & 6, 1964.* Princeton, N. J.: Author, 1964.

FLANAGAN, J. C. Some considerations in the development of situational tests. *Personnel Psychology,* 1954, 7, 461–464.

GRANT, D. L. Situational tests in the assessment of managers. Part II: Contributions to the assessment process. *The Executive Study Conference: Management games in selection and development.* Princeton, N. J.: Educational Testing Service, 1964. Pp. 129–134.

LAURENT, H. *A summary of early identification of management potential.* Research Project, Social Science Research Division of the Employee Relations Department, Standard Oil Co., 1961. (Unpublished report)

LAURENT, H. The early identification of management talent. *The Conference on Executive Study: Identifying management talent.* Princeton, N. J.: Educational Testing Service, 1962.

LOPEZ, F. M. *Evaluating executive decision making: The in-basket technique.* New York: American Management Association, 1966.

MAHONEY, T. A., JERDEE, T. H., & NASH, A. N. *The identification of management potential—A research approach to management development.* Dubuque, Iowa: William C. Brown, 1961.

SPITZER, M. E., & MCNAMARA, W. J. A managerial selection study. *Personnel Psychology,* 1964, 17, 19–40.

TAFT, R. Multiple methods of personality assessment. *Psychological Bulletin,* 1959, **56,** 333–352.

WALKER, H., & LEV, J. *Statistical inference.* New York: Holt, 1953.

(Early publication received October 21, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 107 110

EFFECTS OF CHOICE AND SALES MESSAGE ON CUSTOMER-SALESMAN INTERACTION

JOHN U. FARLEY AND ROBERT L. SWINTH[1]

Graduate School of Industrial Administration, Carnegie Institute of Technology

A group of 87 female Ss were presented alternative sales messages for a roll-up yardstick. 1 presentation stressed product attributes and the other stressed a complimentary personal discussion of how the product is consistent with S's role in life. Each S provided scaled evaluations of the product and the salesman after making her choice between the product and a sum of money. Significant and positive differential evaluation of both product and salesman were found for Ss who chose the product rather than money and those who received the presentation stressing product attributes. Evaluation-response and evaluation-presentation interactions were also significant, although sales rates were approximately equal for the 2 sales messages.

Despite the annual expenditure of billions of dollars on salesmen and sales effort, empirical study of the personal interaction in the sales situation—the salesman's stock and trade—is relatively new. Homans' (1958) idea that "interaction between persons is an exchange of goods, material and non-material [p. 597]" appears consistent with much of the professional sales-management literature, although the latter is often not explicit on the subject. Experiments show that social exchange is characteristic of most interpersonal relationships (Back, 1950; Gerard, 1954). For instance, members of a group are more likely to comply with each others' wishes if they feel liked by the others than if they do not. A field study of insurance salesmen showed that similarity of various perceived physical and social attributes of salesman and buyer increased the likelihood of sale (Evans, 1963)—a result directly analogous to similar features which make for successful field interviewing.[2]

Further, the emotional state resulting from interaction of two people can include evaluations of impersonal objects and events (Heider, 1946), and this has been generalized to a balance theory which predicts equilibrium position for such evaluation-evaluator pairs. For example, if two people share an affection for one another, in equilibrium they will share a feeling (positive or negative) toward an object (Cartwright & Harary, 1956). Applied to the sales situation, this theory implies that a customer who purchases should show more positive feeling toward both salesman and product than a customer who does not purchase, presuming the salesman shows a favorable feeling toward the product. If the balance-theory formulation is correct, analysis of the sales situation requires study of attitude toward the salesman as well as the product, especially in view of the common notion that salesmen are by and large evaluated negatively (Levitt, 1965, p. 31).

In this vein, one might think of the salesman as engaged in two activities that contribute to customer action. The first is providing data about the product so that the customer is able to make an informed choice; the second is providing a degree of personal interaction and recognition. In the latter aspect of his work, the salesman often assumes an inferior social role relative to the customer (Leavitt, 1958). Since self-esteem is an important aspect of human needs, it follows that a basis for exchange may just as well be filling needs for esteem as needs for data. In Homans' (1961) terms, the interaction may "apply, for instance, to the exchange of intangible services for social esteem in a market that is far from perfect [p. 12]." While studies of personal reactions of salespeople to customers are available (Lombard, 1955; Whyte, 1948), much less has been

[1] The authors are indebted to Victor Vroom for constructive comments on the study.

[2] The entire September 1956 issue of *The American Sociologist* dealt with interview-respondent interactions on a number of levels and is germaine to this discussion.

TABLE 1

ANALYSIS OF COVARIANCE OF SCALED RESPONSES TO QUESTIONS ABOUT THE SALES SITUATION

Effects	*df*	*MS*	*F*
Regression on covariate	1	34.4377	12.6681*
Main effects			
Questions (Q)	9	121.1379	44.5611*
Choice (C)	1	64.2505	23.6348*
Pitch (P)	1	67.9098	24.9809*
Interactions			
$Q \times C$	9	10.1087	3.7185*
$Q \times P$	9	8.1949	3.0145*
$P \times C$	1	1.6635	0.6119
$P \times Q \times C$	9	1.7703	0.6512
Adjusted error	745	2.7185	

* $p < .01$.

done in terms of prospects' reactions to salesmen.

This study explores two questions: (*a*) are the implications of the balance theory borne out in the customer-salesman interaction, and (*b*) which is more valuable to the customer—information about the product, or personal recognition that builds his self-esteem? Previous research indicates only that liking both salesman and product is important, but not whether the appeal to one or the other link of exchange is likely to be more effective.

METHOD

Subjects

There were 32 undergraduate female students interviewed individually in a special school room, plus 34 student wives and 21 suburban housewives interviewed at home—in all, a sample of 87 respondents.

Procedure

Each subject (*S*), after hearing a sales presentation which lasted approximately $1\frac{1}{2}$ minutes, chose between a flexible roll-up yardstick and a sum of money. The yardstick was chosen after a pretest of a number of items revealed that its market value was, on the average, perceived correctly and that no pretest *S* owned a close substitute. The money offered was 75¢, the median estimated value from the pretest.

Two presentations were used in an attempt to vary the social exchange between salesman and customer—one which emphasized a description and demonstration of various features of the product, and another which stressed why *S*'s manifest role (student, wife) is important and why the product is consistent with that role. These messages were dubbed, respectively, the "product pitch" and the "personal pitch." The product pitch was identical for all *S*s. The personal pitch was structurally identical although the appeals were composed specially for each of the three groups of *S*s. Each *S* received one message, and the interviewers were given randomized tables giving the explicit order in which the two sales pitches were to be given. Subsequent to choosing between the item and money, each *S* filled in a questionnaire measuring various aspects of her perception of the value of the yardstick, her response to the salesman, and her enjoyment of the interviewing situation.

RESULTS AND DISCUSSION

Analysis of the questionnaire provides implications for both the equilibrium salesman-buyer relationship and the salesman-buyer-product evaluation process. The scaled responses to the 10 questions were analyzed in an analysis of covariance. The covariate was a three-valued discrete variable testing for an intergroup effect for the three groups of *S*s—students, student wives, and housewives. The large *F* value for the regression (Table 1) shows that there is such an effect. The data, adjusted for the covariate, were analyzed in a factorial framework, with the main effects: a question differential; *S*s' decisions at the conclusion of the sales message, that is, whether they chose the money or the product; and the particular sales pitches to which they were exposed. The factorial design permits analysis of the interactions as well as the main effects. A routine using unequal cell sizes was employed because a few questions were missing or not responded to on some interviews (although the questions themselves never varied over questionnaires) and because, of course, it was not possible to control the proportion of *S*s choosing the money or the product after the presentation. Since the salesman showed a favorable attitude toward the product and attempted to exhibit a friendly attitude toward the customer, these links are given and only the customer-product and customer-salesman links are under examination.

The results of the factorial analysis are shown in Table 1. Each of the effects is highly significant, indicating important response differentials among the questions. The analysis normalizes out these differences due

TABLE 2

MEAN SCALE VALUES FOR QUESTIONS WITHIN GROUPS RECEIVING THE SAME SALES MESSAGE AND MAKING THE SAME CHOICE

Question	Sales message received		Respondent choice	
	Product pitch	Personal pitch	Took product	Took money
1. The price I would expect to pay for this product in a retail store.	5.57	5.67	5.65	5.48
2. The desirability of this product to me.	7.29	6.44	7.03	6.34
3. The usefulness of this product to me.	7.55	6.71	7.62	6.34
4. The quality of this product in comparison with others similar in price and function.	7.13	6.67	7.07	6.33
5. The price of this product rather high or rather low.[a]	4.66	4.98	4.83	4.77
6. The salesman was rather pleasant or rather unpleasant.	8.20	7.60	7.93	7.77
7. The salesman's "pitch" was rather boring or rather interesting.[a]	4.33	4.55	4.30	4.93
8. The points the salesman made were rather influential or uninfluential in my decision.	6.47	4.73	6.03	4.12
9. The salesman was rather efficient or rather inefficient.	7.76	6.89	7.42	7.07
10. I rather enjoyed or rather disliked participating in this study.	8.23	7.68	8.06	7.58

[a] Scale directions reversed.

to unequal mean scale values. More meaningful is the fact that each of the other main effects—Ss' choices and the message received—has a significant effect on the responses to the questions. Two of the interactions are also highly significant over and above the main effects—the interactive effect of the type of message received on question responses and the interactive effects of choice (money or product) on response. The other pair-wise interaction and the three-way interactions are not significant, but pooling for degrees of freedom is unnecessary because the other F ratios are already highly significant.

The analysis of covariance establishes that certain effects are present, but within-cell means must be examined to study the nature of the effects. The mean scale values of Ss' responses to the 10 questions for the two main effects of interest (sales message received and respondent choice) are shown in Table 2. Mean values are for an 11-point scale from 0 to 10 for each of the questionnaire items. A larger number means a more positive response, except for the two scales (5 and 7) in which the poles were reversed as a control check to see whether respondents were reading the questionnaire carefully. In these cases, large values imply (5) a higher price and (7) a more boring salesman.

The results are generally consistent with the balance theory, since those who took the product evaluated it *and* the salesman higher than those who did not. The first five items in the questionnaire measure aspects of the respondent's liking for the product—the expected price, desirability, usefulness, quality, and relative price of the product to her. On these five scales, those who received the product pitch and those who took the product, in all but one case, felt more positively toward the product than those who received the personal pitch and those who took the money. The last five items measure liking for the salesman. The respondent indicates how pleasant, interesting, influential, and efficient the salesman was and how much she enjoyed participating in the study. On these latter five scales, those who received the product pitch and those who took the product felt more positively toward the salesman, again with but one exception. These results are

TABLE 3

CHOICE RESPONSES TO SALES MESSAGES

Sales message received	Subjects' choices		Total
	Product	Money	
Product	34	11	45
Personal	34	8	42
Total	68	19	87

corroborated in a further breakdown of S response by sales pitch. Among Ss who received the product pitch, those who took the product indicated more liking for the product and the salesman than those who took the money. Similarly, among Ss who received the personal pitch, those who took the product indicated more liking for the product and the salesman than did those who took the money.

The results show that the equilibrium position for product choosers involves a warmer feeling toward the salesman regardless of sales message and are thus consistent with balance theory. However, the causal direction in this relationship is unclear. Since the messages were randomized over Ss, the scale difference can be imputed to experimental effects, but the single-measurement, post-experiment design provides no insight into the important dynamics of the process. Indeed, there is a distinct possibility of joint causation with complex feedback mechanisms, although the theory is not clear on this point.

Exchange theory is useful in interpreting Ss' evaluations of the two components of the sales messages. The sample as a whole perceived differences in the two sales appeals, and Ss who received the product pitch evaluated both product and salesman more favorably. They thus felt that information about the product was more valuable than information about themselves.

The strong interactions (Table 2) imply that the successful sales appeal includes significant nonadditive components of personal commitment by the buyer to the salesman as well as to the value of the product. The well-balanced sales message thus must pay significant attention to both material and nonmaterial aspects of the sales interaction.

The information exchange data based on attitudes might also imply that the product pitch is more effective in terms of inducing customer choice. However, recipients of each message chose the product (Table 3) with approximately the same frequency—75.7% for the product pitch and 80.9% in the case of the personal pitch. These enviably high sales rates are due primarily to the appeal of the product, but are not significantly different from one another by a chi-square contingency table test. Thus, while both balance hypotheses are consistent with both choice and attitude analyses, the exchange hypotheses are not so consistently supported. This is not, of course, the first time that attitudinal data alone have been insufficient to predict consumer choice.

REFERENCES

BACK, K. W. The exertion of influence through social communication. In L. Festinger, K. W. Back, S. Schachter, H. H. Kelly, & J. Thibaut (Eds.), *Theory and experiment in social communication.* Ann Arbor: University of Michigan, 1950. Pp. 21–36.

CARTWRIGHT, D., & HARARY, F. Structural balance: A generalization of Heider's theory. *Psychological Review*, 1956, **5**, 277–292.

EVANS, F. B. Selling as a dyadic relationship—a new approach. *American Behavioral Scientist*, 1963, **6**, 76–79.

GERARD, H. B. The anchorage of opinions in face-to-face groups. *Human Relations*, 1954, **7**, 313–325.

HEIDER, F. Attitudes and cognitive organization. *Journal of Psychology*, 1946, **1**, 107–112.

HOMANS, G. C. Human behavior as exchange. *American Journal of Sociology*, 1958, **63**, 597–606.

HOMANS, G. C. *Social behavior: Its elementary forms.* New York: Harcourt, Brace & World, 1961.

LEAVITT, H. J. Sales management: Time for a change? *Dun's Review*, April 1958, **71**, 43–45.

LEVITT, T. *Industrial purchasing behavior.* Boston: Harvard University Press, 1965.

LOMBARD, G. F. F. *Behavior in a selling group.* Boston: Harvard University Press, 1955.

WHYTE, W. F. *Human relations in the restaurant industry.* New York: McGraw-Hill, 1948.

(Received March 30, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 111–119

CLINICAL, PSYCHOMETRIC, AND WORK-SAMPLE APPROACHES IN THE PREDICTION OF SUCCESS IN PEACE CORPS TRAINING

LEONARD V. GORDON [1]

Department of Educational Psychology, State University of New York at Albany

A study to determine whether failure to be selected for overseas assignment could be predicted prior to Peace Corps training. 178 Peace Corps volunteers, assigned to 3 programs, reported prior to training for a full week of assessment, which included individual situational tests, paper-and-pencil personality tests, projective tests, and work-sample language training. Predictions of nonselection were made by clinical methods, from language scores alone, from a combination of paper-and-pencil personality test scores alone, and by assessment based on these last 2 sets of scores. All 4 sets of predictions had significant validity. However, no significant differences occurred among methods, the simplest and cheapest approach being as effective as the most complex and costly.

Clinical assessment and psychometric procedures are contrasted by Cronbach (1960) as follows:

> The psychometric tradition isolates separate dimensions of ability and personality and represents the individual by assigning scores on these dimensions . . . [p. 578]. The principal features of assessment procedures are: use of a variety of techniques, primary reliance on observations in unstructured situations, and integration of information by experienced psychologists . . . rather than combining separate scores in a statistical formula [p. 582].

Other differences associated with the two methods include the manner in which the data are gathered and the amount of data gathered and utilized. Clinical assessment procedures typically employ individual rather than group testing methods, tend to require considerably more time for data gathering, and usually are concerned with more data than are psychometric methods.

Despite the greater amount and complexity of data gathered in the assessment approach, there is no evidence to suggest that it is more valid than the psychometric approach. In fact, Meehl's (1954) review of 20 studies, in which predictions by clinical methods were compared with those made by the psychometric approach, found the latter to be far superior. In a later review of the validity of global predictions, Cronbach (1960) concluded that the most ambitious efforts at assessment by psychologists give a generally "black picture." Vernon (1964) in reaching similar conclusions, states

> There seems to be little proof that, by adopting depth conceptions of personality, it becomes possible to improve greatly on our naive method of making predictions about people. . . . Indeed, clinical diagnosis seems to be most effective when the clinician's methods approach more closely those of the statistically trained psychologist . . . [p. 89].

Milholland (1964) in his review of the assessment area provided no new data that would challenge Cronbach's or Vernon's conclusions.

The disappointing results obtained by clinical assessment programs may be due, in part, to their being theory oriented rather than criterion (or problem) oriented. Cronbach (1960) noted that the few successful assessment procedures tended to be performance tests which were essentially "work samples" of the criterion task.

The large amount of data that is typically obtained for each subject (*S*) in these programs also may have contributed to the poor results shown by clinical assessment. Even where the assessor had collected criterion-oriented information, the utility of this in-

[1] The author was temporarily associated with the Peace Corps from December 1964 through February 1965 for purposes of the present study. The author wishes to acknowledge the fine cooperation of many people both within the Peace Corps and elsewhere. Particular appreciation is extended to Robert E. Krug and Allen Kulakow of the Peace Corps, Claude J. Bartlett of the University of Maryland, and the two language coordinators, Guido Capponi and Paul V. Cooper.

formation may have been reduced by the addition of information in which he tended to place an undue amount of faith. Vernon (1964) also made the point that

> the clinician collects such complex evidence that his weighing or synthesis of it is liable to go awry. Having little information on just which predictors are most significant he may put more trust in the poorer than in the better ones [p. 86].

Bartlett and Green (1966) have demonstrated empirically that beyond a few variables the validity of clinical prediction tends to be lowered by the availability of additional data.

Although the weight of evidence, so far available, has not supported the use of clinical assessment for selection programs, and despite its very high cost, such programs continue to be developed. Undoubtedly, the face validity of the problems used and the confidence shown by the assessors contribute materially to the support of such programs.

Even where clinical assessment has demonstrated some validity, this alone does not justify the application of this approach in the operational setting. It is important to ask the question, "Will other, simpler techniques work as well?" Any study which employs expensive and elaborate testing procedures should routinely incorporate and evaluate simpler procedures in order to provide a fair and more complete estimate of the utility of the former.

The present paper describes the results of such an evaluation. The effectiveness of an extensive clinical assessment program is compared with that of two simple approaches, one based on work-sample measurement and the other on psychometric instrumentation. In addition, the lengthy clinical assessment is compared with a "brief assessment" based only on information from the work sample and the psychometric measures.

Operational Objectives of the Research

Under current Peace Corps selection procedures, an applicant submits to the Peace Corps a form on which he provides information such as college grades, work experience, foreign language ability, special skills, etc. At a testing center, he is administered group tests of mental ability, language aptitude, and language achievement (if appropriate). At Peace Corps headquarters an appraisal regarding his suitability for overseas service is derived from these materials and from 6 to 12 letters of reference. This information in conjunction with the applicant's preferences and the availability of a suitable program is used to determine whether to accept him for Peace Corps training.

The Peace Corps training program represents a primary screening device. Here the volunteer is seen for the first time by Peace Corps personnel. On the basis of his performance and other information gathered during training, a Selection Board, consisting of training personnel, a psychologist, a psychiatrist, and a Peace Corps representative, recommends that he be assigned overseas or that he be "selected out" of the Peace Corps. While the factors on which the recommendations are made vary from Board to Board, special skills, performance in language training, psychological evaluation, and peer ratings are typically given some weight by all Boards.

On the average, about 25% of the trainees are selected out or drop out by the end of training. This constitutes a considerable hardship, psychologically and economically, for most of these volunteers. Some have given up their jobs, sold their automobiles and other possessions, and said good-bye to their friends and family for what was to have been a 2-year overseas assignment. In addition, this attrition is costly to the Peace Corps, and frequently prevents the filling of quotas for important programs.

For both economic and compassionate reasons, the Peace Corps undertook a study to determine whether, by means of pretraining assessment, valid predictions could be made as to whether an applicant would successfully complete Peace Corps training. If the results were sufficiently promising, it was intended that a pretraining assessment program would be established operationally for use with all applicants.

For purposes of the present study, being selected for overseas assignment at the end of training was the criterion of interest. While performance overseas is the ultimate criterion, it was not the problem of immediate concern since the deselection rate overseas is low.

having been variously estimated from 3% (Kelley, 1963) to 8% (Shriver, 1965). Also, because of budgetary considerations, the obtaining of follow-up criterion measures was not planned. Because of the urgency of the present research, predictions were to be made for actions of the Selection Boards, as they normally occur. No attempt was to be made to develop uniform procedures across Boards or to refine the Board procedures in any way for purposes of this particular study. It was to be assumed that the Board decisions had acceptable validity.

DEVELOPMENT OF THE RESEARCH PROGRAM

The writer was invited by the Peace Corps to serve as Project Coordinator for the pretraining assessment program, which was to be conducted under contract by Princeton University. Four psychologists had agreed to serve as Project Team and accepted responsibility for the general development and conduct of the assessment program.[2] The Project Coordinator was to provide guidance and assistance to this Project Team in the structuring of the program, and was to appraise the operational utility of the outcome. The Project Team was to submit a set of predictions to the Peace Corps 6 weeks after the termination of the assessment program, and the Project Coordinator was invited to submit an independent set of predictions prior to the maturity of the criterion.

No real choice of Peace Corps programs for use in the present study was possible. The three programs selected were the only ones having a reasonable number of trainees and scheduled to start at the designated time. The *Ss* consisted of applicants who had been accepted for training for programs in Peru, El Salvador, and Thailand. They had agreed to report 1 week early to participate in the assessment program; they were to leave for training at the University of New Mexico, Northern Illinois University, Hilo, Hawaii, or Puerto Rico. At that point, *Ss* were to be informed that the program had been conducted for research purposes and that none of the data would be used in selection. The staff at the training sites knew that the volunteers were to undergo pretraining assessment. However, they had no access to information regarding the nature of the assessment program, and were asked to conduct their training and Board proceedings in the accustomed manner.

Assessment testing was to be performed at two separate locations, a laboratory site at Princeton University and a semiprimitive field site near Tucson, Arizona. The two sites were selected to provide an indication of the relative effectiveness of the two types of settings for operational assessment. The samples directed to each pretraining assessment site had been selected so as to be equivalent with respect to sex, educational level, and destination. There were 54 males and 36 females who reported to Princeton and 50 males and 38 females who reported to Tucson.[3] Two members of the Project Team and equivalent junior staff, trained as a unit at Princeton, served at each site. The assessment program at both sites was the same, except for two group activities which were adapted to their particular settings, and the language programs instituted at each site. The schedules were equally demanding, lasting from about 6:30 in the morning to about 11:30 at night over a full 6-day period. The *Ss* were organized upon arrival into groups of about 18, homogeneous as to sex, for purposes of housing, test scheduling, peer ratings, and group activities.

The assessment program was devoted largely to instrumentation developed by the Project Team, and primarily included semiprojective tests, psychological scaling procedures, and "complex simulated assessment environments." The last consisted of four, specially devised, situational performance tasks, briefly described as follows.

The first task presented *S* with the problem of building an infirmary on the South Sea Island of Wabowa. He could ask as many questions as he wished regarding the island and its cultures, and was to present a tentative solution to the problem of getting the natives' approval and cooperation in the project. In the second, *S* was to describe the system of checks and balances in American government to a "Venezuelan immigrant" role player, who during the course of the explanation revealed a lack of comprehension of particular points. In the third, *S* was to assume that he was conversing with a person from another culture who had strong anti-American attitudes, and was to argue both sides in the imagined discussion. In the fourth task, he was to enlist the aid of an "Indian Government official" in establishing a poultry-raising project, the role-player "official" questioning both *S*'s motives in coming to India and his sensitivity to the Indian culture.

The group activities at the field site consisted of a rabbit cookout and the construction of a wooden privy by each group of *Ss*. At the laboratory site the movie "Mondo Cane" was shown and discussed and a community development proposal was prepared. The *Ss* were rated by observers during the course of these activities.

Since the trainee's performance in Peace Corps language training is typically given some weight by the final Board, a brief work-sample language program was introduced into the pretraining assessment program by the Project Coordinator.[4]

[2] The Project Team consisted of Harold M. Schroder and Bertram L. Koslin of Princeton University, O. J. Harvey of the University of Colorado, and David E. Hunt of Syracuse University.

[3] Inequality of sample size between sites was due to a small number of "no shows."

[4] The two language programs were developed and conducted under subcontract. Guido Capponi, of the

In order to obtain an unbiased estimate of language learning ability, the original specifications called for the use of a single language that would be unfamiliar to all trainees. However, the language department at the University of Arizona, which was to provide training at the Tucson site, did not have sufficient instructional staff for such a language program. Accordingly, a Spanish program was developed for all trainees who had not had more than 1 year of high school Spanish, and a Russian program was established for those who had. The programs were conducted at the University, over a period of 5 nights, for 3 hours per night by university language staff. The Foreign Service Institute (FSI) technique of oral language training, utilized in the Peace Corps language training program, was employed. At the end of the program, rankings, based on language proficiency, were provided by the language staff for all trainees in each language program.

A Vietnamese language program had been prepared in Washington for use at Princeton. Ten days before the initiation of assessment, higher authority decreed that a different language be substituted. A Mandarin Chinese program was rapidly developed, seven Chinese "instructors" (only two had previous teaching experience) were recruited and were provided an average of 6-days training in the FSI oral language teaching method and a lecture on rating procedures. The Princeton language program was conducted over a 3-day period, 3 hours per day. Trainees were taught in groups of six, and were assigned to completely new groups and new instructors for each of the 3 hours. Trainees were ranked by the instructor at the end of each hour, and the sum of the rankings for the last 6 hours served as the student's language score.

In addition to the above instrumentation, a number of paper-and-pencil tests were included in the assessment program. Four of these tests served to provide the psychometric predictions.

Types of Prediction

Clinical assessment. Data from the situational performance tasks and other instrumentation developed by the Project Team, together with such additional data as information obtained from the application form, the mental ability and language-aptitude test scores used in initial selection, and sixth-day peer ratings and work-sample language scores obtained during the assessment program, were combined by the Project Team to provide for each trainee a score on each of the following five basic personality components: (*a*) *Adaptability* represented the individual's conceptual level or level of information processing (Harvey, Hunt, & Schroder, 1961). Lower levels are associated with fewer perspectives and fewer rules for interpreting these perspectives. Adaptability was measured by two paper-and-pencil tests and two situational performance tasks. (*b*) *Motivation* reflected *S*'s commitment to Peace Corps goals as well as the amount of interest shown, his readiness to contribute and learn, and his willingness to take on responsibility and leadership, and was assessed in all group activities and all situational performance tasks. (*c*) *Interpersonal Relations* reflected the trainee's ability to engage in effective and constructive interpersonal relations as judged by observers in the group activities, the four situational performance tasks, and sixth-day peer ratings. (*d*) *Intellectual Level* was measured by scores on the Peace Corps' measure of general mental ability (used in original selection) and college grades. (*e*) *Language Capacity* was measured by the Modern Language Aptitude Test scores (used in Peace Corps selection) and the work-sample language training scores.

Through use of clinical assessment procedures, rather than fixed decision rules, "overall individual predictions were made by the [Project Team] acting much like a Peace Corps assessment board, following a careful consideration of the five component scores [Schroder, Harvey, Hunt, & Koslin, 1965, p. 25]."

The Peace Corps had provided the Project Team with the estimated distribution of Board ratings (on a scale of 1 through 6) for this particular sample of volunteers, and an estimated deselection rate (excluding medical cases) of 16%. This information was used as a guide by the Project Team in making its numerical ratings, and in the identification of 18% of the trainees in the present sample who were judged to be least likely to complete training.

Work-sample measurement. Instructors' final rankings were used as the language scores for the trainees at the field site. Since about twice as many trainees had taken Spanish as had taken Russian, the rank scores for the latter group were simply multiplied by two in order to provide equivalent scores across languages. For the Chinese language program, the sum of the rankings for the last 6 hours provided the most reliable measurement and was used as the language score for the Princeton trainees.

It was not possible to merge Tucson and Princeton language scores because of the very large number of tied ranks and the more restricted range of scores in the latter sample (rankings had been made in groups of six trainees). Accordingly, the validity of the work-sample approach could be evaluated by site, but not by sex across sites, and identification of the 20% of trainees least likely to be recommended for overseas assignment was performed separately by language.

Psychometric measurement. The only psychometric measures considered in the present analysis consisted of scores on four, brief, factored personality

University of Arizona, was responsible for the Tucson program and Paul V. Cooper, of English Language Services, Inc., for the Princeton program. The Project Coordinator provided the specifications for the program, assisted in the development of the Princeton program, and evaluated the predictive effectiveness of both.

tests. Two of the tests measured personality traits (Gordon, 1953, 1956) and two measured values (Gordon, 1960, 1966a).[5] All had been administered under standard instructions and under the assumption, on the part of the examinees, that the scores would be considered in selection. Since there were not sufficiently large samples of the same sex and from the same site to permit the use of cross-validation techniques, a priori decision rules for the integration of the test information had to be established.

On the basis of a review of the available research and impressionistic literature relating to performance of the Peace Corps volunteer, certain personality and value dimensions were identified as being potentially effective predictors of success in the Peace Corps program. These dimensions were assigned simple differential weights reflecting their estimated relative importance. Then, for each of these dimensions, cut points were established at about the tenth and twentieth percentiles (male and female college norms),[6] and 1 point was given for scores between about the twentieth and eleventh percentiles, and 2 points for scores at about the tenth percentile or below. The four tests were then scored for each trainee by multiplying the points obtained on each scale by the weight of that scale and then summating across the scales.[7] Since the scoring procedure had been devised specifically for the identification of individuals in the lower part of the criterion range, the resultant distribution had an extreme negative skew, with 80% of the trainees having a score of 1 or less. The 20% having the higher scores (1.5–15.5) were identified as being the least likely to be selected for overseas assignment.

Brief assessment. Predictions were made by the Project Coordinator utilizing only the language work-sample scores and the scores derived from the 1-hour paper-and-pencil test battery in order to permit a comparison of the validity of assessments made on the basis of a limited amount of information with that based on more extensive data. The language scores and the psychometric scores were combined in a semimechanical fashion, providing a brief assessment score for each *S* in the sample. The language score alone was used in assigning predictor scores to individuals having a psychometric score of zero. For those individuals whose psychometric score was in the lower decile, language was given virtually no weight. Both language and psychometric scores contributed to the final prediction for individuals having intermediate psychometric scores. In assigning predictor scores, somewhat greater significance was attached to the Chinese language program for trainees to be assigned to Thailand, and to the Spanish program for individuals to be assigned to Latin America. The expected distribution provided by the Peace Corps was used as a guide in making ratings (1 through 6), and 18% of the trainees who were judged to be most likely to be deselected were also identified.

RESULTS

Results are presented for all trainees assessed at the laboratory site and for all but five females and one male assessed at the field site. None of the six completed the paper-and-pencil test battery, and three did not complete language training.[8]

As previously indicated, the criterion was simply that of having been selected by the Board for overseas assignment versus not having been selected by the Board. The nonselected group included all volunteers who had resigned during training, who were dropped for medical reasons, and who were deselected by the final Board. All nonselected cases were included in the lower criterion group since deselection for any cause is equally important operationally, representing both a hardship for the individual and a loss for the Peace Corps. Also, the basis of nonselection is not always clearly categorizable. Some volunteers who resign would not have been selected by the Board had they remained, and in some instances medical deselection appears

[5] It would have been desirable to have included other variables in the present analysis. However, all data other than the four personality tests and the work-sample language ratings were processed at Princeton and were not made available for this purpose.

[6] Integral scores do not occur at the tenth or twentieth percentile points in the published norms for some of the scales. In these instances, adjacent percentile values (e.g., the twenty-second percentile) were used as cut points.

[7] The derived scores (of 1 or 2) were multiplied by the following weights in order to reflect the estimated differential importance of the dimensions considered: two for Emotional Stability, Responsibility, and Personal Relations; one for Self-Esteem and Benevolence; and one-half for Support, Vigor, Variety, and Ascendency and Sociability (combined). (Cut points at the eightieth and ninetieth percentiles were used for Support.) Also one-half of a point was added to the individual's score if extreme scores were made on two or more of the following dimensions of possible relevance: Cautiousness, Practical Mindedness, Orderliness, and Recognition. This general approach had been used previously by the author with some success.

[8] Included in the present analyses are all cases for whom psychometric predictor data were available. Omission of the six cases with missing predictor data but for whom clinical assessments were made in no way biased the results with respect to clinical assessment. Conclusions would have been the same had these been included.

TABLE 1

TETRACHORIC VALIDITIES OF THE PREDICTIONS OF SELECTION AND NONSELECTION, BY SITE AND BY SEX, AND PERCENTAGE CORRECT OF PREDICTED NONSELECTIONS FOR EACH PREDICTOR

Predictor	Site		Sex		Total	N'[a]	Percentage correct
	Lab	Field	Male	Female			
Clinical assessment	.38*	.43*	.46*	.28	.39**	31	58.1
Work–sample measurement	.33*	.51**	.42*	.41	.41**	34	58.8
Psychometric measurement	.55**	.14	.38*	.35	.37**	34	55.9
Brief assessment	.51**	.24	.41*	.35	.39**	31	58.1
N	90	82	103	69	172		

[a] N' = Number of deselections predicted.
* $p < .05$.
** $p < .01$.

to have been based on reasons other than physical disqualification.

The respective attrition rates at the Tucson and Princeton sites were 39.0 and 32.2%, with 34.0% of the males and 37.7% of the females falling into the nonselected category. The overall rate of 35.5% is somewhat higher than both the Peace Corps estimate for this sample and the typical attrition rate.[9]

Table 1 provides an indication of the relative effectiveness of clinical assessment, work-sample measurement, psychometric measurement, and brief assessment in predicting nonselection. The prediction of this particular criterion was of great importance to the Peace Corps and represented the primary objective toward which the present research was directed. The tetrachoric correlation coefficients between the dichotomous predictions and criterion of selection versus nonselection are statistically significant for the total sample for all four types of predictors. Also, all predictors are of equivalent validity, the coefficients ranging between .37 and .41. Table 1 also presents the percentage of correct predictions of nonsuccess made by each of the predictor methods. Again, for the total sample all methods yield results significantly better than would be expected by chance, and all results are of similar magnitude. It may be noted that the psychometric measures are differentially effective at the laboratory and field sites. These differences may very well have been a function of differences in conditions of testing.[10] While none of the validities for the female sample are statistically sig-

[9] The loss rate for the entire sample of 178 cases was 35.6%.

[10] During the administration of the personality test battery at the field site, seating arrangements were such that the volunteers did not have adequate privacy of response. This could have accounted for the smaller number of "poor" personality test scores for the Tucson trainees as compared with those at Princeton (despite a higher attrition rate), as well as the lower validity of the psychometric and brief assessment measures for the Tucson site.

TABLE 2

PRODUCT–MOMENT CORRELATIONS AMONG CLINICAL ASSESSMENT, WORK SAMPLE, PSYCHOMETRIC, AND BRIEF ASSESSMENT SCORES, BY ASSESSMENT SITE

Predictor	Laboratory ($N = 90$)				Field ($N = 82$)			
	C	W	P	B	C	W	P	B
Clinical assessment (C)		.25	.07	.35		.57	-.02	.36
Work sample (W)			.13	.57			.08	.52
Psychometric (P)				.83				.69
Brief assessment (B)								

nificant, they are not too different in magnitude from those obtained for the males.

Table 2 presents, by site, product-moment correlations among the four sets of predictors. The relationships between clinical assessment and brief assessment are quite low (.35 and .36). Of the specific individuals predicted as being least likely to be selected by the Board, 9 were the same and 22 were unique on each list for these two types of assessment. It may be noted that less weight is given to the Princeton language program than to the Tucson program in clinical assessment (.25 and .57, respectively), while brief assessment weighted them about equally (.57 and .52). Clinical assessment and psychometric measurement are quite independent (.07 and −.02), while brief assessment and psychometric measurement are highly related (.83 and .69), as would be expected, since the former was based on the latter.

Scores reflecting estimated probability of success for the entire range of *Ss*, using the Peace Corps scale (1 through 6), had been prepared for only the clinical assessment and the brief assessment procedures. Validities, in the form of biserial correlations, of these clinical and brief assessment scores with the dichotomous criterion are presented separately for each sex in Table 3. Clinical assessment has significant validity for the males only and brief assessment has significant validity for the females only. It may also be noted in this table that the two sets of predictor scores have almost identical means and standard deviations, each set of scores having been based on the same Peace Corps guidelines.

TABLE 3

MEANS, STANDARD DEVIATIONS, AND BISERIAL VALIDITIES OF CLINICAL AND BRIEF ASSESSMENT SCORES, BY SEX

Predictor	Male ($N=103$)			Female ($N=69$)		
	M	SD	r_{bis}	M	SD	r_{bis}
Clinical assessment	4.49	.80	.40*	4.40	.78	.22
Brief assessment	4.42	.80	.19	4.58	.79	.40*

* $p < .01$.

Table 4 provides an indication of the relative predictive effectiveness of the three language work-sample programs, for each language group as a whole, and by area of intended assignment. Biserial correlations between the language scores and the dichotomous criterion are presented together with the percentage of correct predictions of nonsuccess. The lowest 20% in each language group were considered to be the poorer prospects. It will be noted that on the whole the Russian program is significantly more valid than either of the other programs, both in terms of prediction for all students in the program and in identifying nonsuccessful trainees.[11] The Spanish program has its

[11] No precise test is available for determining the significance of differences between biserial correlations. By the conventional test for significance of differences between product-moment correlations, differences in favor of the Russian language program are all significant except that of the Spanish language program for the Latin American group.

TABLE 4

BISERIAL VALIDITIES OF WORK-SAMPLE SCORES, AND PERCENTAGE CORRECT OF PREDICTED NONSELECTIONS, BY AREA OF ASSIGNMENT

Language	Latin America				Thailand				Total			
	N[a]	r_{bis}	N'[b]	Percentage	N	r_{bis}	N'	Percentage	N	r_{bis}	N'	Percentage
Chinese	35	.26	11	45.5	55	.26	7	57.1	90	.33*	18	50.0
Spanish	21	.57*	8	75.0	34	.26	3	00.0	55	.40*	11	54.5
Russian	12	.77*	3	100.0	15	.98*	2	100.0	27	.87**	5	100.0

[a] N = Total number of cases.
[b] N' = Number of deselections predicted.
* $p < .05$.
** $p < .01$.

greatest effectiveness for trainees in the Latin American programs (Spanish language training was given in these programs). Chinese has significant validity only for the group as a whole, although it is slightly (but not significantly) better in identifying nonsuccessful volunteers in the Thailand program than in the Latin American program. Present results suggest that while a number of area-specific work-sample language programs might be quite effective in predicting Peace Corps training success (e.g., Spanish), a single non-area specific language (e.g., Russian) may be equally or more effective, and would be considerably more convenient to administer.

Discussion

Taking the results at face value, certain conclusions are suggested on the basis of the present research. First, there do not appear to be any significant differences in the validity of predictions favoring theory-oriented clinical assessment over work-sample measurement, psychometric measurement, or an abbreviated assessment approach. Second, each of these approaches provides prediction significantly greater than would be expected by chance. Third, the level of prediction is of practical utility, and as applied could result in a reduction of the attrition rate by about 5%. Fourth, where relative cost is considered, clinical assessment is by far the poorest choice, and cannot be given serious consideration as an operational alternative. Finally, valid prediction may be effected for the lower part of the criterion range by very simple instrumentation.

Data provided by the 1-hour group-administered personality test battery yielded predictions of nonsuccess not significantly different from those obtained through the use of the far more elaborate and time-consuming measures. These findings agree with those reported by Mischel (1965), who found significant predictive validities for self-report measures but not for global evaluations, against a criterion of staff ratings of overseas performance for a sample of Peace Corps volunteeers. However, unlike the present study, the self-report instruments had been administered to *Ss* under an experimental set. Similarly, Kelly and Fiske (1951), in their research on the selection of candidates for training in clinical psychology, reported that validities of certain standard groups tests were as high as or higher than those for the more clinical procedures, and that the more intensive techniques, in which the judges felt the greatest confidence, worked less well than the apparently more superficial. Many people do respond conscientiously to paper-and-pencil personality tests, particularly in circumstances that promote honest self-appraisal. In the present sample probably not all trainees were convinced that they wanted to go overseas as Peace Corps volunteers, and this, in itself, would tend to encourage frank or honest responses.[12]

The clinical assessment and psychometric predictions were based largely on instrumentation and procedures that the principals had themselves developed, and with which they were thoroughly familiar, thus providing a fair test of the relative effectiveness of these techniques in predicting the designated criterion. Work-sample measurement, on the other hand, was less optimally represented. The hastily contrived Chinese language program, taught at the Princeton site by a nonprofessional staff, placed this work-sample measure at a disadvantage and probably resulted in an underestimate of the validity of this type of predictor. On the basis of present findings, it would appear that a carefully devised work-sample language program in an unfamiliar language might be expected to yield higher validities than either type of assessment incorporated in the present study. The Chinese program did give significant prediction, and the Russian program was strikingly effective.

A number of studies have shown that, beyond some minimum, the use of additional data to make predictions does not help and may actually result in a reduction in validity. There is some evidence that the large amount of information entering into the predictions made by the Project Team actually may have been handicapping. Being unable to process all of their data before the deadline for the

[12] Data for the Peace Corps volunteers on the value tests are summarized in Gordon (1966b).

forwarding their predictions, the Project Team, a short while later, developed a second set of clinical predictions which incorporated all of their information. Schroder et al. (1965) report,

> The accuracy of [the first and second set of] predictions were approximately the same. Evidently, the addition of a number of measures for arriving at component scores . . . did not give us better predictions. Actually, adding these scales slightly reduced the overall discriminating power [p. 31].

In the same regard, a comparison between the validities of the clinical assessment and the brief assessment approach is of interest. The clinical predictions were based on substantially more data than the latter and required the assimilation of five score components as compared with two for the brief assessment, yet the predictive effectiveness of the two types of assessment was quite similar.

A serious deficiency in the use of clinical assessment procedures is the absence of set decision rules that can be applied by other assessors. The Project Team in its report (Schroder et al., 1965) states,

> Since no general rule for combining components is adequate for all individuals, overall scores will be more meaningful if they are generated by a small, trained committee with full knowledge of the overseas assignment for each trainee and the freedom to change the rules of component combination in taking individual factors into account [p. 41].

Since the validity of the predictions is a function of the particular individuals making them, it cannot be assumed that instrumentation found to be valid when used by one "committee" will have any validity when applied by another. The use of a new assessment team, or any change in team composition, would necessitate validation anew. This problem is minimal in psychometric procedures, where fairly rigid decision rules can be established, and in work-sample measurement, where the test scores speak for themselves.

The following quotation from Cronbach (1960) summarizes the implications of the present findings particularly well:

> [Psychometric and impressionistic assesments] are suited to different purposes. When clinical testers answer questions for which their methods and theory are badly suited, their answers are next to worthless or at best are costly beyond their value. . . . Assessment methods have earned a bad name for themselves by trying to compete with measurement techniques on their own ground. In the absence of excellent research to guide the combination of information, the wide band technique should not be advanced as a means of predicting specific, recurring criteria [p. 606].

REFERENCES

BARTLETT, C. J., & GREEN, C. G. Clinical prediction: Does one sometimes know too much? *Journal of Counseling Psychology*, 1966, **13**, 267–270.

CRONBACH, L. J. *Essentials of psychological testing.* New York: Harper, 1960.

GORDON, L. V. *Gordon personal profile.* New York: Harcourt, Brace & World, 1953.

GORDON, L. V. *Gordon personal inventory.* New York: Harcourt, Brace & World, 1956.

GORDON, L. V. *Survey of interpersonal values.* Chicago: Science Research Associates, 1960.

GORDON, L. V. *Survey of personal values.* Chicago: Science Research Associates, 1966. (a)

GORDON, L. V. Values of the Peace Corps volunteer. *Psychological Reports*, 1966, **18**, 328. (b)

HARVEY, O. J., HUNT, D. E., & SCHRODER, H. M. *Conceptual systems and personality organization.* New York: Wiley, 1961.

KELLY, E. L. The future of selection in the peace corps. In *Peace Corps and the behavioral sciences.* Washington: United States Department of Health, Education and Welfare, 1963. Pp. 736–741.

KELLY, E. L., & FISKE, D. W. *Prediction of performance in clinical psychology.* Ann Arbor: University of Michigan Press, 1951.

MEEHL, P. E. *Clinical versus statistical prediction; a theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press, 1954.

MILHOLLAND, J. E. Theory and techniques of assessment. In P. R. Farnsworth, O. McNemar, & Q. McNemar (Ed.), *Annual review of psychology.* Palo Alto: Annual Reviews, 1964. Pp. 331–346.

MISCHEL, W. Predicting the success of Peace Corps volunteers in Nigeria. *Journal of Personality and Social Psychology*, 1965, **1**, 510–516.

SCHRODER, H. M., HARVEY, O. J., HUNT, D. E., & KOSLIN, B. *Component assessment in Peace Corps trainees.* Princeton: Princeton University Press, 1965. (Mimeo)

SHRIVER, R. S. Testimony before the Senate Subcommittee on Constitutional Rights of the Committee on the Judiciary. *American Psychologist*, 1965, **20**, 915–922.

VERNON, P. E. *Personality assessment: A critical survey.* London: Wiley, 1964.

(Received March 31, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 120-130

PERFORMANCE GOALS AS DETERMINANTS OF LEVEL OF PERFORMANCE AND BOREDOM [1]

EDWIN A. LOCKE AND JUDITH F. BRYAN

American Institutes for Research, Washington Office

6 experiments (2 pilot studies and 4 main experiments) are reported dealing with the relationship of performance goals to level of performance and degree of boredom or interest in the task. Tasks used included simple addition, perceptual speed, and psychomotor coordination. Trial times ranged from 2 min. to 2 hr. In the 2 pilot studies postexperimental goal descriptions were significantly related to performance level, and Ss indicated that trying for a specific goal or score was the major source of task interest. In the 4 main experiments a specific hard goal led to a higher level of performance and more task interest than a goal of "do your best." There was no consistent relationship between changes in boredom or interest and changes in performance within the experimental groups.

In spite of much anecdotal evidence indicating that the worker's production goal (intention, quota, or norm) is a major determinant of his level of production (e.g., Mathewson, 1931; Roethlisberger & Dickson, 1939; Smith, 1953; Whyte, 1955), the experimental study of the effects of performance goals on performance level has been given very little attention by psychologists. Thus far there have been a few studies which indicate that level of performance is linearly related to the difficulty (or level) of the goal. Studies by Dey and Kaur (1965) using a letter-cancellation task; Locke (1966c) using a "brainstorming" task; Locke (1966a) in a reanalysis of data gathered by Fryer (1964) using a code-learning task; and Siegal and Fouraker (1960) using an experimental bargaining task, found (generally) that a "harder" goal led to a higher performance level than an "easier" goal. In Locke's (1966c) study, even when the empirical probability of reaching the harder goal was less than 10%, this goal produced higher output than did goals which were easier to reach.

Even less attention has been given to the effects of qualitatively different goals. Mace (1935), in a series of studies using a computation task, found in one experiment that a goal of "do your best" led to a higher level of performance than a goal to surpass a score representative of *S*'s previous performance or than a goal of doing a fixed number of computations. In a second experiment with a similar task he found that giving *S* a specific goal each day based on his initial ability and on the performance of a control group yielded a higher performance level than a goal of "do your best." An experiment by Locke and Bryan (1966) using a complex psychomotor coordination task found that Ss given a specific hard goal on each trial performed better than a matched group of Ss told to do their best.

The first major purpose of the present studies was to further explore the effects of qualitatively different goals on performance utilizing a wider variety of goals than was used previously. In addition, efforts were made to control for differential knowledge of results among the different goal groups (in the study by Locke & Bryan, 1966, the group with hard goals also had knowledge of score, but the do-best group did not). Third, the generality of goal effects across different tasks was of interest. Finally, the lengths of the trials used in the present studies were both longer and shorter than those used previously by Mace (20 minutes); thus the effect of goals as a function of time span could be explored.

A second major purpose of the present studies was to examine the effects of qualita-

[1] This research was supported by Contract Nonr 4792(00) between the Office of Naval Research and the American Institutes for Research. The opinions expressed are not necessarily those of the Department of the Navy.

tively different goals on degree of boredom and interest in the task. Again anecdotal evidence is all there is to go on. In 1934, Wyatt, Frost, and Stock wrote of a group of factory workers whom they had studied intensively:

> One of the most frequent devices adopted by the operatives for the purpose of relieving boredom was the creation of definite aims. From time to time they attempted to complete a certain number of units in a given period; arranged the material in the form of pleasing designs; or competed against each other during short outbursts of great activity [p. 54].

Roy (in Whyte, 1955, pp. 28ff) describes the prevalence of goal-setting activity among workers on piece-rate payment systems and the interest-inducing effects of having definite goals to work for.

The only experimental study on this topic to the author's knowledge was a pilot study by Locke (1966b) of 22 *Ss* working for 5 hours on a psychomotor task. There was evidence that specific goals yielded somewhat more task interest than a goal of "do best," but the results did not reach acceptable levels of significance. It was hypothesized in the present case that working toward a determinate goal would lead to a higher level of task interest than would be the case with an "abstract" goal such as "do your best."

The senior author served as the experimenter (*E*) in the second pilot study and the fourth main experiment, and the junior author served as *E* in the remaining studies. The *Ss* were run in groups in the two pilot studies and in the first main experiment, and individually in the last three experiments.

PILOT STUDY 1

The first pilot study involved 47 male and 25 female paid volunteers recruited from five Washington, D. C., area colleges. The *Ss* worked for four 30-second practice trials and ten 2-minute experimental trials on a number-crossing task described by Moran and Mefferd (1959) as a measure of perceptual speed. The *S*'s task was to cross out each number in a row like the number circled at the beginning of a row of random numbers.

The original experimental conditions were different *goal instructions* given to different groups of *Ss*. These ranged from instructions to "do your best," to "work at a comfortable pace," to "try to beat on each trial a score that is 8 more than your total score on the 4 practice trials." An *F* test on the performance scores of the different experimental groups showed no effect of the goal instructions on performance. However, responses to postexperimental goal checklists indicated that less than 24% of the *Ss* were pursuing their assigned goals. Therefore *Ss* were reclassified according to these postexperimental goal descriptions. Reclassification yielded the following five goal categories.

High-standard group ($N = 9$). This group was composed entirely of *Ss* who indicated that they were trying to reach or exceed the goal of eight more correct rows than they had completed on the four practice trials. These *Ss* actually reached or exceeded their goal on 5% of the trials.

High-improvement group ($N = 16$). The *Ss* in this group stated that they were trying to beat their best previous score by one or more rows on each trial. These *Ss* reached or exceeded their goals on 21% of the trials.

Low-improvement group ($N = 12$). This group consisted of *Ss* who were either trying to equal their best previous score ($N = 4$), trying to beat their immediately previous score ($N = 6$), or those *Ss* trying to equal their immediately previous score ($N = 2$). These groups all had in common the fact that their improvement goals were less difficult than those of the high-improvement *Ss*. These *Ss* were able to reach or exceed their goals 39% of the time.

Do-best group ($N = 22$). The largest a posteriori group was composed of *Ss* who indicated they were trying to do their best without regard to their assigned standard or to their previous scores.

Miscellaneous low-motivation group ($N = 9$). The *Ss* who stated they were working with "little effort," "just trying to complete the task," or "working mainly for accuracy," etc., were placed in this group. These *Ss* had in common the fact that they appeared not to be highly motivated and had no specific (quantitative) goal in mind.

Of the original 72 *Ss*, 4 whose goals did not fit in any of the above categories were dropped from the analysis at this point.

The relative mean improvement scores (over *Ss'* practice scores) of each a posteriori goal subgroup were reasonably consistent across experimental conditions. Thus it seemed permissible to combine the goal subgroups across conditions.

An *F* test performed on the mean improvement scores of the five groups (analysis of covariance) yielded an *F* ratio of 7.28 ($p < .001$). The order of improvement was: (1) High Standard; (2) High Improvement; (3) Do Best; (4) Miscellaneous Low Motivation; and (5) Low Improvement. The *t* ratios which were greater than 2.20 for individual comparisons of means were: each of the first three versus the last group and the first versus the fourth.

This study suggested that performance goals could be used to account for differences in level of performance among groups. It also indicated the importance of determining whether or not *Ss* are

actually trying for the assigned goals (i.e., following instructions) before drawing conclusions about the effects of experimental conditions.

A bipolar graphic rating scale for boredom-interest was given to each *S* after each experimental trial. Analysis of the mean interest scores for the ratings on 10 trials combined did not yield any substantial differences between the a posteriori goal groups. However, when asked to list their reasons for being interested in the task, over 56% of the *S*s gave "trying for a specific goal or standard" or "trying to improve over my previous score" as a reason. This was by far the most frequently given category of response. Also *S*s who listed "trying for a particular goal or score" as a source of task interest were significantly more interested in the task on the 10 experimental trials than were those who did not list it ($t = 2.97$, $p < .01$).

PILOT STUDY 2

The second pilot study used the same *S*s as in the first study, but the task in this experiment was simple addition modeled on Test N-1 from French, Ekstrom, and Price's (1963) reference kit for cognitive factors.

Again, all *S*s were given four 30-second practice trials followed by ten 2-minute experimental trials. As in the previous study the experimental conditions were the different goal instructions given to different groups of *S*s. These ranged from instructions to "beat a score of 8 more than your total practice score," to "do your best," to "beat a score of 6 less than your total practice trial score."

Again, no significant effects of experimental condition on performance were found, but it was found from postexperimental goal checklists that less than 42% of the *S*s were pursuing their assigned goals.

Reclassification of *S*s according to their goal descriptions, however, yielded far more homogeneous performance groups. The order of mean improvement over total practice trial scores for these groups was: (1) *S*s trying to beat their practice scores by more than 8; these *S*s reached or beat this goal less than 4% of the time ($N = 8$); (2) *S*s trying to beat their practice scores, but by less than 8; these *S*s reached or beat this goal 50% of the time ($N = 26$); (3) *S*s trying to do their best ($N = 21$); (4) *S*s trying to equal or get to within 5 points of their practice scores; these *S*s reached or beat their goal 76% of the time ($N = 13$); and (5) *S*s trying to get to within 6 points of their practice scores or worse; these *S*s reached or beat this goal 62% of the time ($N = 4$). Highly significant t ratios for individual comparisons were found between Groups 1 and 2 versus 4 and 5, and 3 versus 4 and 5 combined.

Like the previous study these results suggested that performance goals could be used to account for differences in level of performance among groups and that it was important to determine whether or not *S*s were in fact pursuing the goals they were assigned.

A single boredom-interest questionnaire given at the end of the experiment did not differentiate significantly among the a posteriori goal groups. However, over 61% of the *S*s gave "working toward a goal or standard" or the "challenge of improvement" as a cause of interest in the task. Again, this was by far the most frequently mentioned source of task interest.

The most obvious problem with these two studies, of course, has to do with the validity of the postexperimental goal descriptions. It would not be easy to defend the argument that these goal descriptions were *entirely* post hoc rationalizations of *S*'s performance. First of all, the post hoc goal of "do best" would not really explain anything. Second, some *S*s (those trying for the highest goals in each experiment) checked goals that they reached less than 5% of the time, which would hardly suggest a rationalization of *actual* behavior. Finally, previous findings (Locke, 1966c) that hard goals (experimentally manipulated before performance) lead to higher performance than easy goals are in agreement with some of the present findings. Nevertheless, the possibility remains that *some* *S*s in the previous two experiments were rationalizing their performance rather than describing their actual goals. Thus subsequent studies were done taking more care to get *S*s to try for goals set before the experiment.

It was decided, in the next two studies, to use the same two tasks that were used in the first two studies but to emphasize only two of the goals used previously. The ones chosen were the high-standard or hard goal and the do-best goal. These two goals had *not* produced *significant* differences in performance in the pilot studies, though in both cases the hard-goal *S*s did better than the do-best *S*s. In the next two experiments it was decided to make the experimental trials considerably longer to determine whether the superiority of the hard-goal to do-best *S*s would become significant when the situation allowed more opportunity for motivational (persistence) factors to come into play. The previous two studies had only involved ten 2-minute trials; the two to follow involved single trials between 1½ and 2 hours in length.

Another reason for choosing the specific and do-best goals was that *S*s in the pilot studies indicated that trying for specific goals was the major source of task interest. The alternative to trying for a specific goal is trying for some abstract goal such as "do your best." Even though substantial differences in boredom did not emerge between the hard-goal and do-best groups in the pilot studies, it was felt that they would with longer trial times.

EXPERIMENT 1

METHOD

Subjects

The *S*s were 41 paid, college-student volunteers recruited from the University of Maryland during fall semester. There were 26 male and 15 female *S*s.

Task

The task was perceptual speed as described in Pilot Study 1. This time, however, there were 60 numbered pages of problems instead of 20. To increase ease of scoring, S wrote, in a blank space at the end of each line, the number of digits in each line which matched the circled digit (rather than crossing out the matching digits). The Ss began with two 1-minute practice trials on which they were told to do their best. The 60 pages were in one pile, unstapled, and S put each completed sheet under the pile before beginning the next sheet.

Procedure

The Ss were run in two groups. Both groups worked continuously for 1½ hours at the task, being interrupted only briefly every 15 minutes to make ratings of boredom and to mark the point they had reached in the booklet. The interest-boredom measure was a bipolar, vertical, graphic scale anchored at various points by statements ranging from "it was extremely boring" to "it was fascinating." (The S was asked to describe his degree of task interest or boredom during the previous 15-minute period.) A measure very similar to this one has been found previously to agree well with other measures of boredom for industrial workers (Smith, 1955).

Do-best condition ($N = 17$). The Ss in this group were told to do as many problems as they could in the 1½-hour period.

High-end-goal condition ($N = 24$). On the basis of the performance of the do-best Ss, a table of means was set up showing the obtained 1½-hour scores of the do-best Ss of various initial abilities as indicated by practice trial scores. The actual means shown to the end-goal Ss were the *total attempted* means of the do-best Ss and were about 13% higher than the *total correct* scores for the do-best group. To insure goal acceptance, Ss were offered a $1.00 bonus for beating their "expected" mean score (based on do-best Ss of their initial ability) and an additional bonus of $1.00 for beating the mean score of do-best Ss in the next higher initial ability category. With this information in mind, each S was instructed to indicate in writing what his own goal would be, with the stipulation that it be no lower than the mean score of do-best Ss of his initial ability. (The bonuses were to be paid on the basis of performance, regardless of where the goals were set.) The Ss were reminded that their goals were in terms of number correct and that therefore they should allow some leeway for error in beating their goals. The Ss marked the page in their pile of problem sheets corresponding to their goal with a red X.

All Ss were administered a postexperimental goal questionnaire.

RESULTS

Goal acceptance. Of the 17 do-best Ss, 11 checked the do-best alternative on the postexperimental goal questionnaire. Four additional Ss indicated they were trying for a generally equivalent goal (e.g., "speed"). Two Ss set themselves specific high goals in terms of the number of pages they tried to complete. Of the 24 end-goal Ss, 21 set themselves goals which were harder than the norm sheets indicated they should make on the basis of their initial ability; the remaining 3 set goals appropriate to their initial ability according to the norm sheet. All end-goal Ss indicated on the postexperimental questionnaires that they were trying for their goals (or some score close to it) for part or all of the 90-minute period. Only 2 of the 24 Ss actually reached the goal they set at the beginning of the experiment.

Performance. Figure 1 shows the mean number of rows correct for each goal group for each 15-minute period on the perceptual speed task. It is clear that the end-goal group showed a higher level of performance than the do-best group from the first 15-minute period on, with the size of the difference increasing substantially during the 90-minute period.

The mean improvement in terms of rows correct over the prorated practice scores (on which the two groups were not significantly different) was significantly greater for the end-goal group than for the do-best group ($t = 3.49$, $p < .002$). If the two do-best Ss who indicated they were trying for specific end goals are dropped, the t ratio increases to 4.01 ($p < .001$). The t ratio using linear slope scores was 4.50 ($p < .001$).

On the average the end-goal Ss scored 9% higher than the do-best Ss, the range being −15% to +51%. Of the 24 end-goal Ss, 17 scored higher than do-best Ss of equivalent initial ability, and 2 of the 7 failures were the result of comparisons with the 2 do-best Ss who had set themselves end goals.

Interest-boredom. There was some evidence that specific goals increased interest in the task as compared to the do-best goals. The mean interest score of the end-goal group was greater than that of the do-best group for each of the six 15-minute periods; however, the t ratio for the difference of the means for all ratings combined was only 1.82 ($p < .10$). The slope of the two boredom

reached their subgoals 50% of the time, and six of the nine subgoal Ss reached or exceeded their end goals (i.e., their eighth subgoal).

Performance. Figure 2 shows the mean number of problems correct for each experimental group for each 15-minute period of the 2-hour session. It can be seen that the end-goal group showed the highest level of performance and that its superiority to the do-best group increased during the experiment, reaching a peak after 1½ hours. The subgoal group's performance was between that of the end-goal and do-best groups.

The end-goal group showed a significantly higher mean linear slope in performance than the do-best group (t for matched groups = 2.46, $p < .05$) and got a significantly greater number of problems correct (t for matched groups = 4.50, $p < .001$). Only one end-goal S failed to beat his do-best counterpart; and this difference was only −18 (or −2.5%), whereas some end-goal Ss beat their do-best counterparts by as much as 350 (or +32%), the overall average being 20% more problems correct.

On the whole, the subgoal group did not differ significantly from the end-goal group nor from the do-best group though the latter difference favored the subgoal group. The t ratio for the mean linear slope of the subgoal versus the do-best group was 1.71, and for number of problems correct it was 1.46 (*ns*). (The t ratio for total correct in the middle hour was 1.90, $p < .08$.) A possible explanation for the lesser effectiveness of the subgoals as compared with the end goals is that the subgoals may have actually slowed the pace of some Ss. The subgoal markers yielded feedback about S's progress in relation to his assigned goal; and E observed that some Ss seemed to use the subgoals as limits rather than as minimum guidelines, trying to increase their speed if they were behind, maintaining their pace if not. In fact, six of the nine Ss scored within 10% of (above or below) their do-best counterparts. In contrast, the end-goal Ss did not have such exact knowledge of progress and, in the face of such uncertainty, may have wanted to reach their end goals as soon as possible.

Interest-boredom. The results for the interest-boredom ratings were similar to those for performance except that the end-goal and subgoal groups were even closer together. The end-goal group showed a significantly higher linear slope for interest than the do-best group (t for independent groups = 2.4[illegible], $p < .05$). The corresponding t for the subgoal versus do-best group was 2.02 ($p < .0$[illegible]). Similarly, the mean drop in interest from the first to the last 15-minute period was less for the end-goal and subgoal groups than for the do-best groups (t's for independent groups = 4.61, $p < .001$; and 2.70, $p < .0$[illegible], respectively). Total interest for all ratings combined was significantly greater for the two goal groups (combined) than for the do-best group ($t = 2.12$, $p < .05$).

DISCUSSION

The results of this experiment, where [illegible] monetary incentives were used to insure goal acceptance, were, in terms of the significance of the goal effects, equal or superior to those of the previous experiment. Particularly with respect to boredom-interest, the present results suggested a substantial enhancement of task interest for the specific goal groups as compared to the do-best group.

The final two experiments involved a comparison of the same two goals as in the previous two experiments, but the emphasis here will be on reporting the results for boredom since the results for performance (which completely replicated those obtained previously) are reported elsewhere.

EXPERIMENT 3

The design and performance results of this experiment are reported in detail in Locke and Bryan (1966). Briefly, the task was a complex psychomotor task on which 14 Ss were given specific hard goals (scores) to try for on each of twelve 10-minute trials, and 14 matched Ss were told to do their best on each trial. All Ss were male, paid (for participation only) volunteers from the University of Maryland.

To summarize the performance results briefly, the hard-goal Ss performed at a significantly higher level than the do-best Ss using both linear slope scores and total

scores as criteria. As in the previous two experiments, the difference among the two groups increased as the experiment progressed.

All Ss were administered boredom-interest rating scales, like those used in the previous two studies, after each trial (and before the hard-goal Ss were told whether or not they had beaten their goals). The results are shown in Figure 3.[3]

As in the previous two studies, boredom increased steadily for both groups. In this case the difference in mean boredom-interest ratings of the groups also increased as the experiment went on. The mean total interest score of the hard-goal group was significantly greater than that of the do-best group (t for independent groups $= 3.24$, $p < .01$). The difference in mean overall linear slopes for boredom was not significant, but the mean difference in slopes for Trials 1–3 only was significant at the .01 level. Thus a difference between the two groups in rate of increase in boredom was almost immediate. After Trial 3, boredom increased at about the same rate for both groups.

EXPERIMENT 4

The final experiment was designed to test separately the effects of knowledge of score and the effects of goal setting on boredom. Knowledge of score had been controlled in Experiments 1 and 2 but not 3. The design and effects of the knowledge and goal variables on performance are reported in detail elsewhere (Locke, 1967, in press) and thus will only be summarized here. Briefly the task was simple addition as in Pilot Study 2 and Experiment 2. The design was a 2×2 fixed model. The major variables were knowledge of score versus no knowledge of score and specific hard goals versus do-best goals. There were nine Ss in each of the four cells. Each group of nine Ss was matched with each other group on initial ability on the task and three measures of attitude (liking and interest) toward the task. Each S worked for 5 trials (separated by rest pauses) which alternated between 10 and 15 minutes in length (yielding 1 hour's working time). The

[3] For an explanation of why there were no boredom ratings for Trial 7, see Locke and Bryan, 1966.

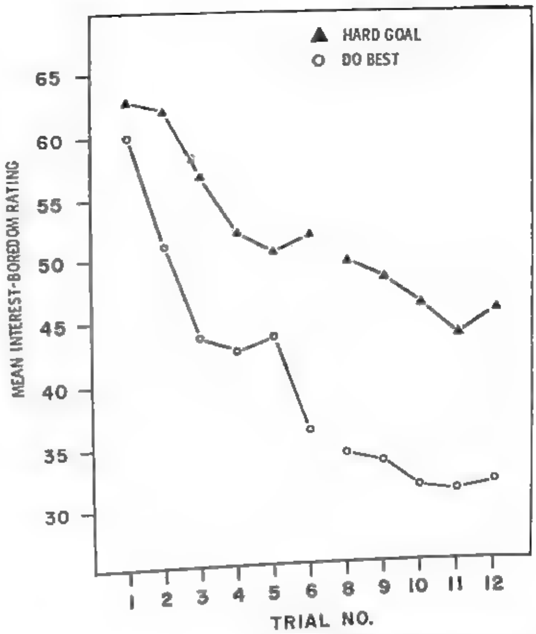


FIG. 3. Mean Interest-Boredom Ratings × Goal Condition (Experiment 3). (Higher scores indicate greater interest.)

Ss wrote their answers on answer sheets each of which they passed to E, through a hole in the wall of the experimental room, after the task was completed. The problems were on separate 3×5 cards put consecutively in a box.[4] The knowledge-of-score Ss were told the number they got right after each trial. The no-knowledge-of-score Ss were not. The do-best Ss were told to do their best on each trial. The hard-goal Ss were given goals before each trial by placing colored index cards vertically at the appropriate point in the problem box. The S was told "you must get to here to reach your goal." The S was not told, however, how many problems this involved. After each trial S was told whether he beat the goal (and was or was not given his actual score depending on his knowledge-of-score condition). The goals of the hard-goal Ss were set by adding about 11% to the scores

[4] The purpose of alternating trial length and having Ss hand in their answer sheets was to prevent do-best and no-knowledge-of-score Ss from getting knowledge of score and from setting specific goals (e.g., such as "improvement") on the basis of their previous scores. The task structure made it very difficult for these Ss to keep track of their progress (e.g., problems correct per minute) or to get any exact idea of their actual scores.

TABLE 1

RELATIONSHIP OF BOREDOM–INTEREST LINEAR SLOPE SCORES TO PERFORMANCE LINEAR SLOPE SCORES IN 4 EXPERIMENTS

Experiment	n	Goal group	r
1	17	Do best	.05[a]
	24	End goal	−.002
2	9	Do best	.58
	18	Subgoal and End goal	−.11
3	14	Do best	−.25
	14	Hard goal	−.07
4	18	Do best	.52*
	18	Hard goal	.55*

[a] A positive correlation indicates the less the decrease in interest, the greater the improvement in performance.
* $p < .05$.

obtained by matched do-best *S*s on corresponding trials. Boredom scales were administered after each trial.

To summarize the performance effects, there was no effect of knowledge of score on performance; but there was a highly significant effect of goals, in favor of the hard-goal condition.

The boredom-interest data were analyzed in several different ways. First, using total boredom scores there was no effect of either variable. However, using linear slope scores there was a significant goal effect in favor of the hard-goal group ($F = 4.33$, $p < .05$). Using the difference between *S*'s boredom rating on the practice trials and his mean rating on the five experimental trials yielded no significant effect; however, using the difference between his practice trial rating and his mean rating on the last two trials yielded a t ratio of 1.99 ($p < .07$) in favor of the hard-goal group. There was no effect of knowledge of score on boredom for any method of analysis. The differential effect of the different goals on boredom did not emerge until the last two trials; on these trials the do-best *S*s increased in boredom considerably while the hard-goal *S*s decreased in boredom (increased interest). The later emergence of the goal effects in this study is probably due to the fact that the trials were shorter than in Experiments 1 and 2. Also in Experiment 3, the boredom difference did not become substantial until after an hour had passed.

Previous research has not given much support to the idea that attitudes and performance are strongly related (e.g., Brayfield & Crockett, 1955), but most previous studies have dealt with satisfaction or liking for the job or task rather than interest and boredom. They have also dealt with static (concurrent) relationships between attitude and performance variables rather than dynamic relationships (i.e., relationships among change scores). In the present studies correlations were computed for each of the four main experiments between linear slope scores for boredom-interest and linear slope scores for performance within each experimental group. The results are shown in Table 1. There was no significant relationship between changes in interest and changes in performance in six of the eight experimental subgroups, and the average correlation for all groups combined was only .18. In terms of group means, it was typical for performance level to increase at the same time that task interest was decreasing, just the opposite of what would be expected if attitude changes were logically related to performance changes. In the present studies *S*'s goals affected both performance and boredom, but the latter were not consistently related to each other.

DISCUSSION

The results of the six experiments reported here strongly support the idea that performance goals are related to and can be used to account for level of performance on a number of different tasks. The results of the pilot studies indicated that postexperimental goal descriptions could be useful in reclassifying *S*s who were not trying for their assigned goals. Experiments 1 and 2 suggest that, with a sufficiently long trial period, *S*s given specific hard goals will perform at a higher level than *S*s trying to do their best. This finding agrees with the performance results of Experiments 3 and 4, which are reported elsewhere. These results also suggest that the financial incentives offered to *S*s for goal attainment in Experiment 1 were not neces-

sary to achieve the results since the same finding was achieved without such incentives in three other experiments. An interesting hypothesis for future research might be that all-or-none incentives which are *contingent upon goal attainment* do not affect performance directly or automatically but function mainly to *commit S to the goal* (providing S values the incentive). It might be worthwhile to compare the effects of an all-or-none incentive (i.e., one given only if a certain score is reached as in Experiment 1) with a piece-rate incentive where S gets paid for each piece or response. One original experimental group in Pilot Study 2 was offered the latter type of incentive, but this group did not perform any better than any of the other original groups.

The present results suggest what might be a more effective means of motivating individuals than telling them to do their best, as is typical in most industrial, educational, and military training and performance situations. Since people appear to respond better to specific than to abstract goals, training time might be shortened and/or productivity raised by inducing Ss to set determinate goals. Fryer (1964) found this to be true on a code-learning task, although Locke (1966a) in a reanalysis of these data found that it was important to control the *level* at which the goals are set.

All four of the main experiments found that giving Ss specific performance goals enhanced (to a greater or lesser degree) interest in the task as compared to Ss with no specific goals (do-best Ss). The effects of goals on interest seem to grow with the length of the experiment and/or the length of the trial periods. These results support the anecdotal evidence of Wyatt et al. (1934) and Roy (in Whyte, 1955) that setting specific goals can function as an antidote to boredom. This has possible implications for industrial situations. It might be possible to reduce boredom on many repetitive tasks by having the workers set determinate goals each day (or hour, etc.). Further research would be needed to determine at what difficulty level goals should ideally be set. Previous experiments (Locke, 1966c) found that the highest productivity was achieved with the hardest goals. On the other hand, Dey and Kaur (1965) did find a slight drop in performance for their hardest goal group (though the significance of this drop was not tested). In addition, Locke (1965, 1966d) found that the hardest goals lead to the lowest degree of satisfaction with, and liking for, the task. Since dislike for the task is liable to lead, eventually, either to apathy or to turnover (Locke, 1966b, found that failure to reach goals makes S less likely to want to perform the task again), some compromise would be necessary to achieve both high output and satisfaction. Roy (in Whyte, 1955, p. 34) suggested that goals of moderate difficulty were maximally effective in getting individuals involved in piece-rate work.

REFERENCES

BRAYFIELD, A. H., & CROCKETT, W. H. Employee attitudes and employee performance. *Psychological Bulletin*, 1955, **52**, 396–424.

DEY, M. K., & KAUR, G. Facilitation of performance by experimentally induced ego motivation. *Journal of General Psychology*, 1965, **73**, 237–247.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. (Eds.) *Kit of reference tests for cognitive factors*. Princeton: Educational Testing Service, 1963.

FRYER, F. W. *An evaluation of level of aspiration as a training procedure*. Englewood Cliffs, N. J.: Prentice Hall, 1964.

LOCKE, E. A. The relationship of task success to task liking and satisfaction. *Journal of Applied Psychology*, 1965, **49**, 379–385.

LOCKE, E. A. A closer look at level of aspiration as a training procedure: A reanalysis of Fryer's data. *Journal of Applied Psychology*, 1966, **50**, 417–422. (a)

LOCKE, E. A. A preliminary study of some cognitive factors in boredom. American Institutes for Research, Washington Office, Report AIR-R&D31-3/66-TR, 1966. (b)

LOCKE, E. A. The relationship of intentions to level of performance. *Journal of Applied Psychology*, 1966, **50**, 60–66. (c)

LOCKE, E. A. The relationship of task success to task liking: A replication. *Psychological Reports*, 1966, **18**, 552–554. (d)

LOCKE, E. A. Motivational effects of knowledge of results: The influence of goal setting. *Journal of Applied Psychology*, 1967, **51**, in press.

LOCKE, E. A., & BRYAN, J. F. Cognitive aspects of psychomotor performance: The effects of performance goals on level of performance. *Journal of Applied Psychology*, 1966, **50**, 286–291.

MACE, C. A. Incentives: Some experimental studies. Industrial Health Research Board (Great Britain), 1935, Report No. 72.

MATHEWSON, S. B. *Restriction of output among unorganized workers.* New York: Viking Press, 1931.

MORAN, L. J., & MEFFERD, R. B. Repetitive psychometric measures. *Psychological Reports,* 1959, **5**, 269–275.

ROETHLISBERGER, F. J., & DICKSON, W. J. *Management and the worker.* Cambridge: Harvard University Press, 1939.

SIEGAL, S., & FOURAKER, L. E. *Bargaining and group decision making.* New York: McGraw-Hill, 1960. Pp. 61–69.

SMITH, P. C. The curve of output as a criterion of boredom. *Journal of Applied Psychology,* 1953, **37**, 69–74.

SMITH, P. C. The prediction of individual differences in susceptibility to industrial monotony. *Journal of Applied Psychology,* 1955, **39**, 322–329.

WHYTE, W. F. *Money and motivation.* New York: Wiley, 1955. Pp. 28–38.

WYATT, S., FROST, L., & STOCK, F. G. L. Incentives in repetitive work. Industrial Health Research Board (Great Britain), 1934, Report No. 69.

(Received April 6, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 131–138

PREDICTION OF CREATIVITY AND OTHER PERFORMANCE MEASURES FROM BIOGRAPHICAL INFORMATION AMONG PHARMACEUTICAL SCIENTISTS [1]

MICHAEL F. TUCKER, VICTOR B. CLINE
University of Utah

AND JAMES R. SCHMITT
Richardson-Merrell, Inc., New York City

A 160-item Biographical Inventory (BI) was administered to 157 pharmaceutical scientists who were randomly assigned to 2 groups of 79 and 78 Ss each. Each alternative from each item in the BI was correlated with 19 criterion measures obtained on each of the scientists. These primarily involved ratings of creativity, quantity of work produced, skill with people, etc., obtained from supervisors, peers, and subordinates. There existed remarkably little relationship between supervisory ratings and peer ratings of the scientists on most variables. Using a double cross-validation design, BI predictor keys were developed for each of the criterion ratings and applied across to the new independent sample. Significant cross validities were obtained, notably in the prediction of the creativity criterion, where correlations of .36 and .42 resulted across the 2 subsamples.

The identification and prediction of scientific competence, especially creative scientific competence, is a difficult and hazardous undertaking. However, it is quite apparent that the future of our progress as a society and a nation depends to a great extent upon our creative scientific talent in all fields. Recently, definite progress has been made in the identification and prediction of such talent through the use of biographical information as predictor data.

Some of the more recent studies demonstrating the value of biographical information in the identification and prediction of scientific competence and creativity are those of Kulberg and Owens (1960), Albright and Glennon (1961), Smith, Albright, and Glennon (1961), Chambers (1964), McDermid (1965), and Taylor, Ellison, and Tucker (1965).

As indicated in these studies, the overwhelming majority of past research in the area has been in engineering and the physical sciences. As a result, there are at present extensive biographical descriptions of the more creative and competent people in these fields. However, with the exception of Roe (1951, 1953), and a few others, little of this research has been concerned with other types of professional people, such as those in the biological or pharmaceutical sciences.

This study was undertaken to determine whether biographical characteristics also could be found and utilized effectively for the prediction of performance criteria of pharmaceutical scientists. Since pharmaceutical scientists engaged in research and development have a potential widespread influence on the general public through the products they develop, it is crucial that the best scientific talent available be selected to fill these positions. A secondary purpose of this study was the development of a valid biographical inventory which, hopefully, would aid in the identification and selection of such high-level scientific talent.

[1] The authors would like to express great appreciation to Frank Sessions, Dan Anderson, and Loretta Tucker for their considerable assistance in collecting and analyzing the data. In addition they would also like to make note of the invaluable assistance rendered by the management and personnel of Richardson-Merrell, Inc., without whose assistance this study could never have been accomplished.

METHOD

Subjects

There were 234 employees of the Richardson-Merrell, Inc., pharmaceutical company who participated in the study. They were employed in the scientific division of Vicks Chemical, the Merrell Drug Company, and the National Drug Company at the time of the study. Through the elimination of females (who were too few in number) and all those classified as technicians (whose work was so routine as to preclude opportunities for innovation, creativity, etc.), as well as those upon whom there was incomplete data, a total of 157 scientists remained in the sample, all of whom had degrees ranging from BA or BS through PhD or MD.

TABLE 1

LIST OF CRITERION MEASURES OBTAINED ON 157 PHARMACEUTICAL SCIENTISTS

Ratings	Variable
Supervisory ratings	1. Quantity of work produced. 2. Skill in getting along with people. 3. Creativity. 4. Overall work performance.
Peer ratings	5. Quantity of work produced. 6. Skill in getting along with people. 7. Creativity. 8. Overall work performance.
Combined supervisory, peer, and subordinate ratings	9. Quantity of work produced. 10. Skill in getting along with people. 11. Creativity. 12. Overall work performance.
Criterion measures derived from company records	13. Tenure I (months employed by company). 14. No. salary increases.
Rankings as provided by supervisors and peers	15. Rank (supervisory). 16. Rank (peer).
Supervisory ratings from a checklist instrument	17. Creativity checklist. 18. Likability.
Performance rating provided by research team	19. Psychologist's rating.
Control measures not included in the BI predictions	20. Tenure II (months under present supervisor). 21. Percentage of time spent working with supervisor. 22. Degree of confidence expressed by a rater in providing a performance appraisal.

The following presents a breakdown of the: (*a*) ages, (*b*) academic degrees obtained, and (*c*) tenure with the company for the remaining sample.

(*a*) Age breakdown:
- 12 people were under 26 years of age.
- 48 people were 27 to 34 years of age.
- 48 people were 35 to 42 years of age.
- 35 people were 43 to 50 years of age.
- 14 people were over 50 years of age.
- Mean age = 37 years.

(*b*) Breakdown of academic degrees obtained:
- 42 people had received a BA or BS degree.
- 35 people had received the Bachelor's degree and had some graduate work, but no graduate degree.
- 9 people had received an MA or MS degree.
- 12 people had some graduate work beyond the Master's degree.
- 59 people had the PhD or MD degree.

(*c*) Tenure information:
- The minimum time with the company was 3 months.
- The maximum time with the company was 33 years.
- The average time with the company was 89 months.

In relation to each other, these research scientists occupied either supervisory, peer, or subordinate positions, so that most of the hierarchical structure was represented.

Data Collection

A 160-item Biographical Inventory (BI) was administered to each participating scientist. This inventory had been developed by Taylor, Ellison, and Tucker (1965) in their extensive studies with NASA scientists. The items found to be most valid and reliable in the prediction of various performance criteria in these NASA studies were retained and assembled into the present form of the BI. Other items also were included which were believed on rational, logical, or empirical grounds to be related to competence and creativity.

At the same time the BIs were being administered, criterion data were collected on each scientist from on-the-job performance ratings by their supervisors, peers, and, in some cases, subordinates. Table 1 presents a list of the criterion measures obtained from these ratings.

With the exception of Variables 20, 21, and 22 (which were used for control and validity purposes) all of the measures presented in Table 1 were later included in the biographical analysis as separate

criterion scores. Variable 20 was used to check the reliability of the supervisory ratings, and Variable 21 was used as part of the Likability criterion. Variable 22 was used as a control measure to check the reliability of all performance appraisals; that is, if a rater was *not* very confident in giving a particular performance rating or appraisal, that appraisal was *not* used in the study.

Procedure

The criterion measures were intercorrelated to determine the extent of overlap among the rating variables across ratees and raters.

An item analysis was then performed upon the BI in relation to each of the criterion measures; that is, a biserial correlation coefficient was computed for each alternative of each of the BI items in relation to each of the criterion measures. By setting minimum validity levels (see footnote in Tables 2 and 3), empirical predictor keys were thus made up on the BI to predict each of the criteria. By splitting the total sample of 157 into two subsamples of 79 and 78 subjects, a double cross-validation was performed.

RESULTS AND DISCUSSION

Tables 2 and 3 give summary statistical results for both samples of scientists. The first 19 variables (reading down the left-hand column) in each table are the criterion measures. On the extreme right are given their means and standard deviations, and in the upper center are given their intercorrelations. It can be noted that all of the performance appraisals provided by supervisors (Variables 1, 2, 3, 4, 15, and 17) show relatively high intercorrelations among themselves in both tables, indicating the influence of a halo effect. This is further demonstrated in the relationships between the Likability criterion (Variable 18) and the supervisory ratings, the former having significant correlations only with these supervisory ratings. The peer ratings (Variables 5, 6, 7, 8, and 16) also shown high intercorrelations, the magnitude of these correlations being generally less than those among the supervisory ratings in Table 2, and generally higher than the supervisory ratings in Table 3. The "total" or average supervisory, peer, and subordinate ratings (Variables 9, 10, 11, and 12) also show significant intercorrelations in both tables, although the magnitude of these correlations is generally much less than those for either supervisory or peer ratings taken alone.

The intercorrelations of the peer ratings on the four basic rating variables (Creativity, Skill with People, Quantity of Work, and Overall Work Performance) with the supervisory ratings on these four variables show that there is very little relationship between peer ratings and supervisory ratings. Evidently when a supervisor rates a man on Overall Work Performance (for example) he is thinking of quite different things than are the peers who rate this same man on this same variable. In examining the pattern of intercorrelations, it would appear that "liking a man" plays a major role in a supervisor giving an employee a high rating on *any* variable. Likability (Variable 18) is significantly related (r's of .41 to .59) to supervisory ratings but *not* to peer ratings. It may be that such things as loyalty, friendship, being pleasant to work with, doing what the boss says, etc., have a significant impact on the supervisor and produce a halo effect which colors all of his ratings. However, the combination of peer, subordinate, and supervisory ratings provides a three-dimensional picture which represents more independent measures of the quantity of work produced, skill with people, creativity, and overall work performance than either the supervisory or peer ratings taken alone.

The Tenure criterion (Variable 13) and the Number of Salary Increases (Variable 14) seem to be generally unrelated to any of the other criterion measures save each other, indicating that length of time on the job and number of monetary increases are not related to the various performance criteria for this sample of scientists. The Psychologist's Rating (Variable 19) shows highest correlations across the two tables with Variable 12, the "total" (or combined) estimate of Overall Work Performance.

In the center section of Table 2 (Variables 20 through 38) the empirical keys constructed from Sample 1 are listed. On the extreme right are listed the means and standard deviations of these keys along with the number of items making up each key. The correlations listed in this section are the result of applying these Sample 1 scoring keys back to the BI responses of the *same* sample (Sample 1) and then correlating the

TABLE 2

CRITERION INTERCORRELATIONS, INITIAL AND CROSS VALIDITIES, NUMBER OF ITEMS PER BIOGRAPHICAL SCORE, MEANS, AND STANDARD DEVIATIONS FOR SAMPLE 1 (79 R-M-I SCIENTISTS)

Variable	No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	No. Items keyed	*M*	*SD*
Sample 1 criteria																							
Quantity of Work (Supervisory)	1	—																			—	96.11	32.77
Skill with People (Supervisory)	2	83	—																		—	88.49	31.80
Creativity Rating (Supervisory	3	79	71	—																	—	83.42	31.07
Overall Work Performance (Supervisory)	4	91	86	83	—																—	95.92	32.69
Quantity of Work (Peer)	5	16	04	13	15	—															—	97.47	24.40
Skill with People (Peer)	6	30	27	26	29	66	—														—	92.96	25.48
Creativity Rating (Peer)	7	16	07	24	15	75	59	—													—	90.72	25.03
Overall Work Performance (Peer)	8	18	09	18	16	94	71	80	—												—	98.53	24.23
Quantity of Work (Total)	9	13	00	08	12	76	44	54	67	—											—	101.77	15.74
Skill with People (Total)	10	28	40	25	34	32	70	30	37	50	—										—	96.95	17.62
Creativity Rating (Total)	11	01	−01	28	07	46	31	59	50	70	45	—									—	95.03	16.22
Overall Work Performance (Total)	12	15	12	19	21	68	44	54	69	90	59	79	—								—	102.66	11.72
Tenure	13	−10	−08	−18	−13	08	04	00	15	02	−04	06	06	—							—	86.34	80.94
No. Salary Increases	14	01	−04	−07	−05	−01	−05	06	06	03	−05	09	06	78	—						—	5.51	5.64
Rank (Supervisory)	15	46	43	45	50	07	21	21	19	10	31	23	23	04	15	—					—	41.47	26.34
Rank (Peer)	16	−04	−08	02	−01	43	43	46	46	34	26	39	35	25	20	29	—				—	52.09	11.85
Creativity Checklist	17	60	53	72	61	−03	−01	06	−01	18	14	32	22	−17	02	43	07	—			—	22.03	10.94
Likability	18	59	53	41	54	−25	−07	−19	−23	−15	11	−22	−13	−20	−08	28	−21	39	—		—	25.27	9.09
Psychologist's Rating	19	09	02	09	11	25	12	26	29	34	17	34	41	05	−03	12	14	02	−07	—	—	9.18	2.86
Sample 1 BI keys																							
Quantity of Work (Supervisory) Key	20	80	67	60	71	18	32	16	18	11	25	01	11	−06	05	37	−01	47	39	−02	54	101.81	7.68
Skill with People (Supervisory) Key	21	71	86	60	75	05	26	10	10	01	41	03	13	−08	−08	39	−07	40	39	−02	62	100.06	7.97
Creativity Rating (Supervisory) Key	22	67	65	85	70	15	21	27	17	09	21	26	16	−12	−06	31	−04	59	30	02	50	102.35	6.51
Overall Work Performance (Supervisory) Key[a]	23	67	67	56	71	15	27	15	19	14	38	13	26	−04	00	44	01	43	36	05	20	102.82	3.33
Quantity of Work (Peer) Key	24	16	01	13	15	88	59	68	83	63	27	37	55	10	02	13	40	−03	−20	20	59	100.70	8.18
Skill with People (Peer) Key	25	24	20	15	22	54	86	43	56	29	57	10	28	00	−02	14	32	−11	−03	−02	68	102.68	9.65
Creativity Rating (Peer) Key[a]	26	13	11	17	14	49	29	72	52	39	19	47	41	−05	−08	16	23	07	−13	19	32	101.46	5.26
Overall Work Performance (Peer) Key	27	15	08	16	15	83	62	75	87	56	33	42	58	21	10	22	45	−04	−25	26	60	101.80	8.21
Quantity of Work (Total) Key	28	04	−05	01	05	58	31	49	51	84	44	61	75	−08	−03	11	26	12	−16	27	65	97.06	9.07
Skill with People (Total) Key	29	15	30	11	23	31	58	29	31	45	85	38	50	−09	−08	23	19	04	06	06	67	99.70	9.78
Creativity Rating (Total) Key[a]	30	02	03	17	06	32	19	48	33	54	34	71	59	−05	02	16	18	20	−24	21	28	102.16	1.90
Overall Work Performance (Total) Key	31	08	07	08	13	58	35	49	58	80	53	64	85	02	06	19	27	13	−15	32	67	97.56	8.74
Tenure Key[a]	32	03	03	−08	−03	10	12	03	17	−05	−01	−09	−04	80	61	09	18	−10	−02	01	28	102.22	5.33
No. Salary Increases Key[a]	33	07	00	−06	−02	−01	03	−02	05	03	03	03	05	64	80	17	13	03	10	00	28	100.47	5.41
Rank (Supervisory) Key	34	31	34	26	36	11	22	25	18	12	29	18	20	03	13	85	32	31	17	08	58	100.72	8.27
Rank (Peer) Key[a]	35	02	−03	05	04	30	37	31	35	21	21	23	23	24	25	28	72	06	−11	20	17	99.79	3.08
Creativity Checklist Key	36	42	37	52	44	−04	−12	06	−06	22	09	33	23	−18	08	28	06	86	29	−03	57	100.42	7.33
Likability Key	37	43	37	30	38	−19	−05	−26	−20	−13	08	−22	−13	−18	−03	16	−18	27	82	−23	54	103.38	7.43
Psychologist's Rating Key	38	−07	−09	−01	−02	19	−03	24	23	28	02	34	35	06	−04	05	08	01	−23	3	63	94.63	9.34

TABLE 2—(*Continued*)

Variable	*No.*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*No. items keyed*	*M*	*SD*
Sample 2 BI keys																							
Quantity of Work (Supervisory) Key	*39*	***03***	*10*	*11*	*04*	*−04*	*06*	*18*	*05*	*−01*	*10*	*22*	*11*	*08*	*08*	*04*	*06*	*12*	*10*	*24*	66	105.15	4.51
Skill with People (Supervisory) Key[a]	*40*	*14*	***30***	*17*	*20*	*−03*	*12*	*10*	*08*	*−09*	*15*	*03*	*05*	*06*	*−07*	*19*	*06*	*10*	*−05*	*13*	24	104.30	3.09
Creativity Rating (Supervisory) Key	*41*	*−09*	*−01*	***06***	*−05*	*06*	*−09*	*29*	*10*	*15*	*−02*	*34*	*19*	*−01*	*−05*	*−07*	*06*	*12*	*−17*	*20*	85	102.43	9.43
Overall Work Performance (Supervisory) Key	*42*	*−01*	*08*	*12*	***01***	*08*	*−02*	*32*	*13*	*06*	*−03*	*25*	*12*	*−01*	*−01*	*−01*	*01*	*15*	*−10*	*21*	69	101.63	6.61
Quantity of Work (Peer) Key	43	13	06	00	11	**13**	16	14	13	16	16	18	17	07	09	20	−02	02	−01	14	62	96.81	4.80
Skill with People (Peer) Key	44	21	19	04	19	08	**34**	05	11	00	33	−13	04	06	14	23	06	−07	11	−02	61	101.58	5.40
Creativity Rating (Peer) Key	45	15	12	09	11	16	12	**22**	12	19	16	15	13	−13	−09	02	−11	07	02	−01	61	100.54	5.87
Overall Work Performance (Peer) Key	46	20	12	07	12	20	30	24	**22**	16	28	16	17	02	04	21	−01	−05	−02	04	64	100.05	4.93
Quantity of Work (Total) Key	47	02	07	05	02	09	15	19	16	**09**	18	26	21	18	15	20	15	00	−14	12	66	107.03	4.96
Skill with People (Total) Key	48	08	15	07	08	01	26	12	08	−09	**20**	−04	00	07	07	19	18	−07	03	−16	67	103.15	4.68
Creativity Rating (Total) Key	49	−01	05	10	03	20	08	37	17	28	16	**36**	26	−19	−16	03	12	11	−05	07	82	102.41	7.73
Overall Work Performance (Total) Key	50	07	13	17	09	22	27	40	31	15	26	34	**31**	03	07	29	18	01	−10	09	73	105.25	5.01
Tenure Key[a]	51	−03	−07	−14	−06	01	04	−08	00	−11	−05	−16	−15	**48**	42	08	21	−16	00	05	29	98.54	3.84
No. Salary Increases Key[a]	52	06	−08	01	00	00	−11	−07	−01	−12	−21	−13	−16	46	**43**	10	07	00	09	−08	22	102.61	2.83
Rank (Supervisory) Key	53	02	05	06	−01	14	07	15	16	20	11	20	20	−01	−07	−03	−07	03	07	26	64	104.39	5.09
Rank (Peer) Key	54	00	−05	−03	−04	26	25	21	22	29	16	13	20	13	14	04	12	−03	−13	20	48	99.37	3.90
Creativity Checklist Key	55	03	10	13	04	10	03	32	15	16	09	32	21	−11	−14	03	01	12	−12	−14	81	110.84	8.66
Likability Key	56	23	15	14	21	18	34	21	25	−03	12	−05	05	15	11	16	17	09	22	11	62	107.20	4.88
Psychologist's Rating Key	57	20	26	26	18	15	16	23	20	09	09	15	20	−01	14	10	10	23	06	07	68	102.77	4.59

Note.—Decimal points omitted: $r_{.05} = .22$; $r_{.01} = .28$.
[a] These keys were constructed on the basis of at least 25% of the sample responding and a minimum correlation of .25 between an item alternative and the appropriate criterion. The other keys were constructed on a 10% and .22 basis.

resulting key-score distributions with the criterion measures in this sample. The same information is presented for Sample 2 in the corresponding section of Table 3. The correlations in heavy print (down the diagonal) each represent the correlation coefficient between a particular key and the criterion it was specifically constructed to predict. The other correlations in this section represent the relationship of each particular predictor key with the *other* criterion measures. These initial correlations down the diagonal are extremely high, ranging from .71 to .88.

Despite these extremely high initial correlations, the acid test of how well the predictor keys work is when they are applied to an entirely new sample. This is done in the bottom section of Tables 2 and 3. In Table 2 the Sample 2 derived keys have been applied to Sample 1; and in Table 3 the Sample 1 keys have been applied to Sample 2. The correlations printed down the diagonal in heavy print in these sections thus represent cross-validity coefficients for a particular key and its corresponding criterion in the new sample.

The examination of these bottom sections in both tables provides a remarkable display of the shrinkage in magnitude of the correlations between empirically constructed predictor keys and criteria when the keys are subjected to cross validation. Whereas all of the BI predictor keys had demonstrated extremely high correlations with the criteria in the initial validating samples, these correlations decreased dramatically in the cross validation.

However, it appears that six variables held up quite well in the double cross-validation. These appear to have significant predictive validity. The variable number, identification, and cross validities of these six variables against their appropriate criteria as presented in Table 2 and Table 3, respectively, are as follows: Variable 47, the Quantity-of-Work key, cross validated .09 and .34; Variable 48, the Skill-with-People key, .20 and .40; Variable 49, the Creativity key, .36 and .42; Variable 50, the Overall-Work-Performance key, .31 and .32; Variable 51, the Tenure key, .48 and .51; and Variable 52, the Num-

TABLE 3

CRITERION INTERCORRELATIONS, INITIAL AND CROSS VALIDITIES, NUMBER OF ITEMS PER BIOGRAPHICAL SCORE, MEANS, AND STANDARD DEVIATIONS FOR SAMPLE 2 (78 R-M-I SCIENTISTS)

Variable	No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	No. items keyed	*M*	*SD*
Sample 2 criteria																							
Quantity of Work (Supervisory)	1	—																			—	96.83	30.00
Skill with People (Supervisory)	2	59	—																		—	86.62	31.19
Creativity Rating (Supervisory)	3	73	51	—																	—	89.92	29.98
Overall Work Performance (Supervisory)	4	83	63	88	—																—	96.18	28.47
Quantity of Work (Peer)	5	25	05	14	14	—															—	93.36	20.01
Skill with People (Peer)	6	06	17	−04	00	72	—														—	91.11	29.88
Creativity Rating (Peer)	7	18	06	24	15	82	71	—													—	90.41	26.32
Overall Work Performance (Peer)	8	25	12	19	17	91	85	90	—												—	95.58	26.82
Quantity of Work (Total)	9	54	33	30	42	63	10	15	56	—											—	97.45	14.97
Skill with People (Total)	10	11	56	04	15	17	58	18	33	38	—										—	95.65	19.84
Creativity Rating (Total)	11	34	29	59	46	35	25	59	47	55	33	—									—	95.53	15.85
Overall Work Performance (Total)	12	40	41	35	43	46	43	13	57	74	65	69	—								—	100.83	11.49
Tenure	13	01	−08	−18	−11	17	03	08	05	00	−07	−22	−02	—							—	90.58	84.74
No. Salary Increases	14	01	−07	−07	−06	22	11	07	17	14	05	06	21	62	—						—	5.90	6.17
Rank (Supervisory)	15	51	31	39	41	18	06	18	18	26	02	22	19	−01	−05	—					—	16.85	23.22
Rank (Peer)	16	15	03	08	03	63	58	58	64	33	20	24	25	05	09	17	—				—	17.99	18.13
Creativity Checklist	17	67	47	81	75	23	08	31	28	41	09	51	36	−21	−18	57	24	—			—	23.11	10.27
Likability	18	53	50	40	53	−15	−01	−11	−07	01	17	15	09	−01	05	30	−01	35	—		—	25.06	9.13
Psychologist's Rating	19	14	12	21	25	01	02	05	07	22	25	20	38	−01	12	08	−08	18	06		—	9.67	2.33
Sample 2 keys																							
Quantity of Work (Supervisory) Key	20	**85**	53	65	65	30	12	22	29	60	14	39	15	01	07	51	18	62	39	15	66	103.63	9.11
Skill with People (Supervisory) Key[a]	21	44	**70**	42	47	05	10	02	09	26	10	24	35	−13	−06	27	06	39	38	01	24	103.26	4.72
Creativity Rating (Supervisory) Key	22	49	39	**80**	61	07	−09	19	11	33	04	53	33	−26	09	33	05	68	11	21	85	102.67	14.10
Overall Work Performance (Supervisory) Key	23	59	48	78	**75**	15	−02	21	19	46	16	56	50	17	−04	41	11	71	22	29	69	102.05	10.26
Quantity of Work (Peer) Key	24	27	15	10	14	**84**	59	69	78	56	17	30	44	20	25	24	54	23	−03	07	62	96.17	8.10
Skill with People (Peer) Key	25	10	15	−12	−03	52	**76**	50	62	30	43	10	29	11	10	09	38	−02	10	−01	61	99.63	9.41
Creativity Rating (Peer) Key	26	19	18	26	19	63	54	**85**	73	36	18	58	11	16	06	31	12	39	−02	12	61	100.56	9.12
Overall Work Performance (Peer) Key	27	24	25	15	19	73	64	74	**81**	50	31	41	55	10	25	28	47	29	00	08	64	99.18	9.12
Quantity of Work (Total) Key	28	57	40	45	46	46	24	31	41	**82**	31	48	66	00	08	10	27	49	16	25	66	106.03	9.96
Skill with People (Total) Key	29	07	48	01	16	06	42	06	18	26	**83**	24	51	−07	09	05	10	03	22	18	67	101.78	9.41
Creativity Rating (Total) Key	30	28	30	57	43	17	07	43	29	36	22	**82**	53	−35	−04	29	16	52	11	22	82	103.45	11.11
Overall Work Performance (Total) Key	31	38	41	39	44	38	27	39	46	64	50	66	**85**	−08	17	31	25	42	10	31	73	104.28	10.13
Tenure Key[a]	32	−06	−16	−30	−21	07	01	−14	−01	−05	−10	−30	−10	**80**	50	−15	04	−31	00	−10	29	99.08	6.46
No. Salary Increases Key[a]	33	−09	−11	−26	−17	15	11	−04	10	08	06	−10	09	73	**74**	−10	10	−26	02	−01	22	102.83	4.44
Rank (Supervisory) Key	34	47	33	46	43	27	07	30	27	40	07	42	34	05	−02	**84**	18	58	16	13	64	103.96	8.93
Rank (Peer) Key	35	25	08	17	11	59	49	50	57	39	20	31	33	09	18	24	**84**	29	03	−01	48	99.32	6.35
Creativity Checklist Key	36	44	37	69	56	17	03	29	22	40	11	55	38	−28	−15	43	17	**80**	17	19	81	110.95	13.85
Likability Key	37	49	49	27	41	00	14	−02	05	24	34	16	17	−03	05	30	10	31	**83**	10	62	105.77	7.70
Psychologist's Rating Key	38	19	14	25	28	17	05	18	19	34	19	34	44	−03	10	17	02	20	05	**84**	68	102.96	8.95

TABLE 3—(*Continued*)

Variable	No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	No. items keyed	*M*	*SD*
Sample 1 keys																							
Quantity of Work (Supervisory) Key	39	06	15	−05	03	07	06	04	07	02	06	−02	02	19	00	01	00	−04	14	07	54	102.10	4.47
Skill with People (Supervisory) Key	40	04	29	10	10	11	07	14	11	02	16	11	08	−01	02	14	09	19	07	02	62	100.47	4.63
Creativity Rating (Supervisory) Key	41	07	18	04	11	01	04	11	09	08	17	21	20	−02	−03	09	−02	14	17	07	50	103.83	4.01
Overall Work Performance (Supervisory) Key[a]	42	17	23	15	12	19	01	17	16	21	06	23	19	19	08	22	09	25	12	09	20	103.62	2.70
Quantity of Work (Peer) Key	43	05	16	06	07	19	14	14	20	18	08	14	14	22	23	09	−03	00	17	09	59	101.81	4.19
Skill with People (Peer) Key	44	07	06	−08	04	17	23	13	21	10	18	05	13	04	11	14	16	−04	16	03	68	102.33	5.54
Creativity Rating (Peer) Key[a]	45	17	24	48	28	11	−02	22	15	15	07	41	22	−15	00	15	04	44	15	15	32	101.60	3.75
Overall Work Performance (Peer) Key	46	14	25	18	18	16	09	08	16	28	17	20	27	16	18	15	02	09	20	12	60	102.39	4.24
Quantity of Work (Total) Key	47	23	21	39	26	26	12	35	31	34	09	48	36	−18	−01	24	05	32	01	29	65	98.65	4.82
Skill with People (Total) Key	48	05	20	−01	02	26	40	29	31	23	40	23	32	−11	−01	20	26	09	03	08	67	100.31	1.91
Creativity Rating (Total) Key[a]	49	33	26	52	39	23	08	27	24	40	09	42	32	−14	00	20	09	49	08	23	28	103.32	3.58
Overall Work Performance (Total) Key	50	16	16	24	16	16	05	19	17	26	10	26	32	−16	02	10	02	14	−01	35	67	99.37	4.31
Tenure Key[a]	51	05	02	−17	−03	02	01	−20	−07	05	02	−26	−03	51	24	−11	−05	−21	08	−02	28	101.97	3.77
No. Salary Increases Key[a]	52	01	−06	−20	−12	11	04	−09	03	08	−04	−17	−02	40	30	−14	−10	−25	02	−06	28	100.33	3.39
Rank (Supervisory) Key	53	10	20	−07	06	25	24	08	20	27	31	06	25	05	14	03	11	−14	16	14	58	100.92	4.79
Rank (Peer) Key[a]	54	15	06	03	05	00	07	−04	01	07	11	02	08	−03	09	−11	18	−04	20	01	17	100.13	2.18
Creativity Checklist Key	55	−02	−04	12	04	−03	−08	11	01	04	−03	23	08	−21	−19	06	−09	12	02	28	57	101.46	3.93
Likability Key	56	02	08	−12	−01	05	06	04	04	−03	08	01	−08	−08	−05	12	−01	02	15	−09	51	103.27	1.84
Psychologist's Rating Key	57	20	07	34	21	06	−05	−01	01	13	−11	04	02	−05	−07	01	−01	14	−02	10	63	95.56	1.92

Note.—Decimal points omitted: $r_{.05} = .22$; $r_{.01} = .28$.

[a] These keys were constructed on the basis of at least 25% of the sample responding and a minimum correlation of .25 between an item alternative and the appropriate criterion. The other keys were constructed on a 10% and .22 basis.

ber-of-Salary-Increases key, .43 and .30. The first four keys, as indicated, are not based on either supervisory ratings alone or peer ratings alone but rather on the *combined* supervisory, peer, and, in some cases, subordinate ratings.

Two other variables also held up quite well in cross validation. These were Variable 40, the key predicting the supervisory Skill-with-People ratings, and Variable 44, the key predicting the peer Skill-with-People ratings. Variable 40 cross validated .30 and .29, respectively, in Tables 2 and 3, and Variable 44 cross validated .34 and .23 as presented in these two tables. Apparently, Skill with People is either an easier attribute to rate or people are generally more aware of this ability in others than they are of the other performances being rated, and therefore the supervisors alone and the peers alone provided more predictable appraisals of this ability than they did with other ratings.

It is important to note that the keys predicting the supervisory ratings alone (with the exception of Variable 40, the Skill-with-People key) did not hold up in cross validation. Also, the keys predicting peer ratings alone, although showing some predictive validity (such as Variable 44, the Skill-with-People key), generally did not hold up too well in cross validation. However, when peer and supervisory ratings are combined, apparently a more accurate estimate of a scientist's performance is provided which can be predicted with some success from biographical information. It is also important to observe here that the performance criteria derived from ranking procedures (Variables 15 and 16) were unpredictable. Although these rank distributions were statistically treated according to Ghiselli and Brown (1962), perhaps the conditions under which they were collected (heterogeneity and large discrepancies among numbers of subjects across ranking groups, etc.) rendered them somewhat inaccurate.

Several important findings with significant implications have emerged from these analyses. Biographical characteristics were identified which, when incorporated into scoring keys, were predictive of several dimensions of the performance of pharmaceutical scien-

tists. This was especially true in the prediction of creativity, where cross validities of .36 and .42 were obtained. It has been shown that peer and supervisory ratings yield relatively low intercorrelations with each other and that either taken alone does not provide as independent or as predictable criterion measures as do combined ratings. This finding is in agreement with and provides some support for the findings of Buckner (1959) who found that

> ratings of shipboard performance for which interrater agreement estimates were high, were less predictable (from scores on two aptitude tests and school achievement) than were ratings for which the interrater agreement estimates were moderate and low [p. 63].

In the present study the interrater agreements between a scientist's supervisors and his peers were extremely low, although the mean rating for a scientist derived from these sources provided quite predictable measures of performance. This finding certainly would appear to be logical in that no one person (or even class of persons) can observe a scientist employee from all different perspectives and view all phases of his performance; whereas several observers (especially at differing hierarchical levels) viewing the same person can get, in sum total, a better overall view of his capabilities.

This study also provides a sobering note to investigators in assessment and applied psychology on the need for and importance of cross validating any empirical keys developed or constructed on any research.

REFERENCES

Albright, L. E., & Glennon, J. R. Personal history correlates of physical scientists' career aspirations. *Journal of Applied Psychology,* 1961, **45,** 281–284.

Buckner, D. N. The predictability of ratings as a function of interrater agreement. *Journal of Applied Psychology,* 1959, **43,** 60–64.

Chambers, J. A. Creative scientists of today. *Science,* 1964, **145,** 1203–1205.

Ghiselli, E. E., & Brown, C. W. Ranking methods. In T. L. Whisler & S. F. Harper (Eds.), *Performance appraisal: Research and practice.* New York: Holt, Rinehart & Winston, 1962. Pp. 203–208.

Kulberg, G. E., & Owens, W. A. Some life history antecedents of engineering interests. *Journal of Educational Psychology,* 1960, **51,** 26–31.

McDermid, C. D. Some correlates of creativity in engineering personnel. *Journal of Applied Psychology,* 1965, **49,** 14–19.

Roe, A. A psychological study of eminent biologists. *Psychological Monographs,* 1951, **65**(14, Whole No. 331).

Roe, A. A psychological study of eminent psychologists and anthropologists and comparison with biological and physical scientists. *Psychological Monographs,* 1953, **67**(2, Whole No. 287).

Smith, W. J., Albright, L. E., & Glennon, J. R. The prediction of research competence and creativity from personal history. *Journal of Applied Psychology,* 1961, **45,** 59–62.

Taylor, C. W., Ellison, R. L., & Tucker, M. F. *Biographical information and the prediction of multiple criteria of success in science.* NASA Research Project, NASA-105, 1965.

(Received April 14, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 139–144

COMPARATIVE STUDY OF NEED SATISFACTIONS IN MILITARY AND BUSINESS HIERARCHIES [1]

LYMAN W. PORTER AND VANCE F. MITCHELL

University of California, Berkeley

703 commissioned officers and 594 noncommissioned personnel serving in an overseas Air Force Command completed a questionnaire measuring need fulfillment and satisfaction. Results for 3 levels of the commissioned officers were compared to previous results for analogous levels of civilian managers. The findings showed that the military officers were less fulfilled and less satisfied than their civilian counterparts. However, fulfillment and satisfaction increased in relation to military rank in the same way as for civilian managers. When commissioned officers were compared with noncommissioned officers, higher NCOs reported more fulfillment but less satisfaction than lower-ranking commissioned officers.

In his book, *The Human Side of Enterprise*, Douglas McGregor (1960) stated that classical principles of organization theory are "derived from inappropriate models [p. 16]." McGregor says:

> The conventional principles were derived primarily from the study of models (the military and the Catholic Church) which differ in important respects from modern industrial organizations . . . . If there are universal principles common to all forms of organization, it is now apparent that they are not the ones derived by classical theorists from the Church and the military [p. 16].

Despite this prominent attention focused on the military model as a key to understanding classical organization theory, psychologists have seldom studied this kind of organization in comparison with business firms (Lange, 1965). The present study is aimed at such a comparison in the area of need fulfillments and need satisfactions.

Recent studies (Porter, 1962, 1964) involving samples of respondents from business enterprises found that managers at higher-level positions reported significantly more need satisfaction than those at lower levels. Theoretical comparisons of military and civilian organizations (Etzioni, 1961; Janowitz, 1959) would lead us to expect a similar relationship within the hierarchy of military managers. These same studies, however, note certain characteristics of the environment within which the peacetime military operates that raise doubts as to whether the amounts of need satisfactions in the military would correspond with those of civilian managers. Military organizations are confronted with the necessity for a sustained high level of preparedness for extreme emergencies (i.e., combat) in a time of nominal peace. In addition, the increasingly shortened reaction time associated with nuclear or even limited warfare emphasizes the need for centralized decision making and continuation of a basically downward-oriented hierarchical structure. At the same time, the sophistication of modern weapons systems and the growing complexity of organizational forms have led to the emergence of more indirect forms of control (Lange, 1965). These disparate requirements—coupled with the fact that most day-to-day military operations are carried out in peaceful, nonemergency situations—point to the strong possibility of relatively greater frustration and dissatisfaction on the part of military officers compared with business managers.

Few studies have been reported that compare satisfactions of military personnel with those in other kinds of organizations. There is some fragmentary evidence, however, in support of our expectations. Fleishman's (1956) study that compared officers in four naval organizations with managerial personnel in business organizations found that the officers believed that they had to assume

[1] The valuable assistance provided by personnel of the United States Air Force is gratefully acknowledged. Particular thanks are due E. V. Baravelle for his assistance in data collection and Tom West for programming and computation.

more responsibility relative to their perceptions of their authority. In another study (Bowers, 1962), a comparison of Air Force research and development staff personnel with staff managers in a civilian research agency showed a greater discrepancy in the military between ratings of the importance of various work factors and ratings of the extent to which these factors were actually provided in the employment situation.

In the current study, perceptions of need fulfillment and satisfaction with fulfillment of commissioned officers from an Air Force Command were compared with those of analogous levels of managers in civilian industries as reported previously by Porter (1962, 1964). In addition, similar attitude data obtained from noncommissioned officers were compared with those of the commissioned officers.

METHOD

Questionnaire

The data for this study were collected by means of the same questionnaire, slightly modified for applicability to military respondents, which was previously used in studies of managers in business firms (Porter, 1961, 1962). The results reported here are based on answers to parts of 13 items contained in the questionnaire (which has been described in detail in the previous articles). These items are relevant to a Maslow-type classification of needs according to their prepotency and thus tap the following five types of needs: security, social, esteem, autonomy, and self-actualization. For each of 13 specific items (e.g., "the feeling of self-esteem") the respondent was asked to indicate, on a 1–7 rating scale: (*a*) How much is there now? (*b*) How much should there be? and (*c*) How important is this to me?

Procedure and Sample

The questionnaire was sent to approximately 800 commissioned officers and 700 noncommissioned officers throughout an Air Force overseas Command.[2] The sample was stratified to obtain proportionate representation from the hierarchical levels of grade and organization and from the different functional activities comprising the Command. Only officers occupying managerial-type positions and noncommissioned officers occupying supervisory-type positions were included. The questionnaire was distributed through the central personnel offices servicing the various bases and headquarters u… and the anonymity of individual respondents w… guaranteed. Usable responses were obtained fr… 1,297 individuals or about 85% of the sample … both commissioned and noncommissioned ranks.

From personal data questions asked on the l… page of the questionnaire, it was possible … classify respondents on a number of independ… variables. The relevant variable for this study w… military grade, since, among commissioned offic… grade correlates highly with level of manageri… responsibility, and among noncommissioned offic… it is the principal determinant of the supervis… hierarchy.

RESULTS

Comparisons between Military and Civili… Respondents

In this part of the Results section, com… parisons will be made between three lev… of the commissioned officers sampled in th… present study and three analogous levels o… managers and executives from business an… industrial firms sampled previously with th… the same questionnaire (Porter, 1962, 1964… In selecting the appropriate comparison lev… from the managerial sample, attention wa… given to relative responsibility for men… money, and material, in addition to the bas… criterion of hierarchical level. The followi… pairings were selected as the most appropria… ones:

Brigadier generals and colonels . . . . Vice-president…
Lieutenant colonels and majors . . . . Upper-middle manage…
Captains and lieutenants . . . . Lower-middle manage…

Table 1 presents the results for need fulfil… ment. The larger an entry in a given cell o… the table, the greater the fulfillment for th… need category for a given group of respon… ents.

It is clear from Table 1 that the variou… levels of the military perceived less nee… fulfillment than equivalent-level managers i… business firms. Signed-rank tests of differenc… of mean values between pairs of ranks for th… 13 individual questionnaire items showe… that upper-middle and lower-middle levels o… managers report significantly more fulfillme… ($p < .02$) than the comparable milita…

[2] The command consisted of an area headquarters and a number of geographically dispersed air bases which were charged with operational flying and a wide variety of logistical activities.

TABLE 1

MEAN FULFILLMENT FOR EACH NEED CATEGORY: AIR FORCE OFFICERS VERSUS CIVILIAN MANAGERS

Need category	B.G./Col. (N = 22)	V.Pres. (N = 611)	L/C/Maj. (N = 217)	Upper middle (N = 659)	Cpt/Lt. (N = 464)	Lower middle (N = 431)
Security	5.77	5.40	4.85	5.17	4.75	5.19
Social	5.64	5.39	5.33	5.19	5.21	5.20
Esteem	5.26	5.49	4.90	5.03	4.52	4.79
Autonomy	5.39	5.65	4.74	5.04	4.48	4.75
Self-actualization	5.30	5.49	4.81	5.13	4.58	4.90
M of 13 items	5.41	5.52	4.89	5.09	4.64	4.90

Note.—Abbreviations: B.G./Col. = Brigadier General/Colonel group; V.Pres. = Vice-President; L/C/Maj. = Lieutenant/Colonel/Major group; Cpt/Lt. = Captain/Lieutenant group.

respondents. Although vice-presidents also reported greater fulfillment than generals and colonels, the differences are not significant by a signed-rank test. Only in the social need area did officers indicate greater fulfillment than managers across all three levels. There were no clear differences between the two samples in the pattern of *relative* fulfillment among the five types of needs.

Table 2 presents the results for mean dissatisfaction, or, in other words, mean deficiencies in need fulfillment. The degree of dissatisfaction was obtained by subtracting the answer to part *a* of each item ("How much of the characteristic is *now connected* with your assignment?") from part *b* of each item ("How much of the characteristic do you think *should be connected* with your assignment?"). Thus, the larger the mean value in a cell in Table 2, the greater the dissatisfaction.

Consistent and significant differences ($p < .01$ by signed-rank tests across 13 individual items) are apparent in the direction of greater dissatisfaction among the military officers than civilian managers of equivalent rank. This conclusion holds up strongly for each of the three comparison pairs of levels, with the most marked differences occurring in the esteem and security need categories.

In terms of ranges in mean dissatisfaction (for the total of all 13 items) from the highest to the lowest ranks in each sample, Table 2 shows that the range for the military (.85–1.21) is about one-third larger than that for the civilian managers (.56–.81). This indicates that the influence of military rank on need satisfaction may be somewhat greater than the effect of echelon level in civilian business organizations.

When relative differences among the five types of needs, in terms of dissatisfaction, are ascertained from Table 2, it can be seen that both managers and military officers indicate

TABLE 2

MEAN DISSATISFACTION FOR EACH NEED CATEGORY: AIR FORCE OFFICERS VERSUS CIVILIAN MANAGERS

Need category	B.G./Col. (N = 22)	V.Pres. (N = 611)	L/C/Maj. (N = 217)	Upper middle (N = 659)	Cpt/Lt. (N = 464)	Lower middle (N = 431)
Security	0.55	0.45	1.14	0.41	1.10	0.38
Social	0.32	0.30	0.61	0.34	0.58	0.33
Esteem	1.04	0.45	1.15	0.66	1.27	0.71
Autonomy	0.90	0.55	1.26	0.87	1.25	0.96
Self-actualization	1.04	0.90	1.45	1.12	1.53	1.17
M of 13 items	.85	.56	1.17	.76	1.21	.81

Note.—For abbreviations see Note of Table 1.

that the most dissatisfied are the self-actualization needs. However, the two samples differ in terms of next-most-dissatisfied need. For the managers it is autonomy, while for the officers it is the esteem need area. Security and social needs are the best satisfied for both samples.

Comparisons within the Military Hierarchy

Table 3 presents the results for need fulfillment for the three officer groups and for the three noncommissioned groups. The table shows a clear trend (significant across the 13 individual questionnaire items at the .01 level by the signed-rank test) within the commissioned officers for each higher grade to report more fulfillment. That is, brigadier generals and colonels consistently perceived more need fulfillment in their positions than did lieutenant colonels and majors, who in turn perceived more than captains and lieutenants. The overall trend within the noncommissioned category is not as clear-cut, since there is relatively little difference in fulfillment between the high-ranking chief master sergeant/senior master sergeant (CMS/SMS) group, and the middle-ranking master sergeant/technical sergeant (MS/TS) group. The latter group does, however, report significantly greater fulfillment than the lowest noncommissioned level in the sample, staff sergeants.

The most interesting result shown in Table 3 is the finding that chief and senior master sergeants reported consistently (and significantly) *greater* fulfillment than the lowest grouping of commissioned officers, the captains and lieutenants. In fact, the top noncommissioned grades (CMS/SMS) reported about as much fulfillment as the lieutenant colonels and majors! Even the middle group of noncommissioned officers (the MS/TS group) reported somewhat more fulfillment than captains and lieutenants. All of this means that there was *not* a steady increase in perceived need fulfillment from the lowest noncommissioned ranks to the highest commissioned officer levels. The increase was discontinuous, with a sharp drop between the highest noncommissioned level and the lowest officer level. The question then arises whether this same break appears with regard to how well the various military levels are satisfied with the fulfillment they receive.

Table 4 presents the results for mean dissatisfaction. These show a significant increase in dissatisfaction from the brigadier general/colonel group to the lieutenant colonel/major group, but virtually no change from the latter to the level of captains and lieutenants. Within the noncommissioned officers, there are strong trends for dissatisfaction to increase from the higher to the lower grades.

Comparing officers with noncommissioned officers in Table 4 shows that senior sergeants (CMS/SMS) reported generally as much or *more dissatisfaction* than captains and lieutenants even though (as was seen in Table 3) they had reported significantly greater fulfillment. Also, the intermediate noncommissioned grades, the master and technical sergeants, who likewise reported more fulfillment than the captains and lieutenants ar

TABLE 3

MEAN FULFILLMENT FOR EACH NEED CATEGORY: AIR FORCE OFFICERS AND NONCOMMISSIONED OFFICERS

Need category	B.G./Col. ($N = 22$)	L/C/Maj. ($N = 217$)	Cpt/Lt. ($N = 464$)	CMS/SMS ($N = 79$)	MS/TS ($N = 302$)	S/S ($N = 2$[illegible])
Security	5.77	4.85	4.75	5.22	4.85	4.43
Social	5.64	5.33	5.21	5.60	5.70	5.48
Esteem	5.26	4.90	4.52	4.92	4.84	4.27
Autonomy	5.39	4.74	4.48	4.83	4.78	4.32
Self-actualization	5.30	4.81	4.58	4.86	4.86	4.48
M of 13 items	5.41	4.89	4.64	5.00	4.96	4.53

Note.—For abbreviations see Note of Table 1. Additional abbreviations are: CMS/SMS = Chief Master Sergeant/Senior Master Sergeant group; MS/TS = Master Sergeant/Technical Sergeant group; S/S = Staff Sergeant group.

TABLE 4

MEAN DISSATISFACTION FOR EACH NEED CATEGORY: AIR FORCE OFFICERS AND NONCOMMISSIONED OFFICERS

Need category	B.G./Col. (N = 22)	L/C/Maj. (N = 217)	Cpt/Lt. (N = 464)	CMS/SMS (N = 79)	MS/TS (N = 302)	S/S (N = 213)
Security	0.55	1.14	1.10	1.20	1.62	1.83
Social	0.32	0.61	0.58	0.56	0.57	0.66
Esteem	1.04	1.15	1.27	1.36	1.45	1.60
Autonomy	0.90	1.26	1.25	1.35	1.46	1.56
Self-actualization	1.04	1.45	1.53	1.54	1.63	1.74
M of 13 items	.85	1.17	1.21	1.26	1.37	1.49

Note.—For abbreviations see Notes of Tables 1 and 3.

about as much fulfillment as majors and lieutenant colonels, are distinctly more dissatisfied than these lower- and middle-officer grades. Thus, the findings shown in Table 4 indicate that although high-level noncommissioned officers felt they were obtaining somewhat more need fulfillment than low-level commissioned officers, at the same time they expected even more fulfillment than the commissioned officers and hence had greater dissatisfaction with their level of fulfillment. This was especially true in the security, esteem, and autonomy need areas.

If comparisons are made among the five types of needs in terms of relative deficiencies, it can be seen from Table 4 that both commissioned and noncommissioned officers tended to regard the self-actualization needs as the most dissatisfied. The second most dissatisfied needs for the officers were esteem needs, but for the enlisted men it was the security need area.

DISCUSSION

In the comparisons between the three ranks of commissioned officers and the three comparable levels of managers, several conclusions seem apparent:

First, military officers tend to be much more dissatisfied at each rank than their civilian managerial counterparts. This finding, though, may not be as important as the fact that the largest differences between the two groups were *not* concentrated in the autonomy and self-actualization categories, as might have been predicted on the basis of the theoretical studies (Etzioni, 1961; Janowitz, 1959) cited earlier. Consequently, the safest generalization at the present time is that although military officers report generally more dissatisfaction than civilian managers, they are not at a particular disadvantage in higher-order need areas.

Second, the hierarchy of commissioned-officer positions in a military organization has roughly the same relationship to perceptions of need fulfillment and satisfaction as hierarchies of managerial positions in business organizations. This was shown by the fact that, as in business and industrial firms (Porter, 1962, 1964), fulfillment and satisfaction increased with increasing rank. More specifically, both samples showed relatively small increases between the lower-middle (i.e., captains/lieutenants in the military hierarchy) and the upper-middle (i.e., lieutenant colonels/majors) levels, and relatively large increases between the upper-middle and vice-president (brigadier general/colonel) levels. Thus, there was a remarkable similarity in the findings as related to differences between adjacent levels.

One of the most intriguing findings emerging from the study was the fact that for perceptions of fulfillment in the military sample there were clearly two sets of hierarchical relationships, one for the commissioned officers and another for the noncommissioned. Since the higher-ranking noncommissioned men consistently reported more fulfillment than the lower-level officers, it appears that each category of respondents (enlisted men and officers) used its own group as a frame of reference in responding to the question-

naire. That is, apparently the high-ranking noncoms, the technical sergeants and above, were responding in terms of the fulfillment they were receiving in their positions or roles as the top-level personnel in a hierarchy going from private to chief master sergeant. The lieutenants and captains, on the other hand, apparently were responding to their roles not as individuals higher in rank than top-level noncoms, but rather as individuals lower in rank than other commissioned officers. This interpretation is reinforced by the fact that not only did the top noncommissioned officers report more fulfillment than the lower commissioned officers, but they also expected more fulfillment. (This was indicated by the fact that their perceived *deficiencies* in fulfillment—"expectations" minus "reality"—were larger than the deficiencies for officers even though their reported fulfillment was greater.) The question then arises as to whether similar findings would be obtained in business organizations if the need perceptions of lead men or quarter men were measured by the same scales as those used with managers. Would the highest-level workers report greater need fulfillment than the lowest-level managers? The available evidence (see Porter & Lawler, 1965) indicates they might not, but there has as yet been no direct test of this question.

REFERENCES

BOWERS, R. V. Members' perceptions and evaluations as measures of organizational effectiveness. In R. V. Bowers (Ed.), *Studies in organizational effectiveness*. Washington, D. C.: Air Force Office of Scientific Research, 1962. Pp. 115–179.

ETZIONI, A. *A comparative analysis of complex organizations*. New York: Free Press of Glencoe, 1961.

FLEISHMAN, E. A. Differences between military and industrial organizations. In R. M. Stogdill & C. L. Shartle (Eds.), Patterns of administrative performance. *Research Monograph No. 81*. Columbus: Ohio State University, Bureau of Business Research, 1956. Pp. 31–38.

JANOWITZ, M. *Sociology and the military establishment*. New York: Russell Sage Foundation, 1959.

LANGE, K. Military organizations. In J. G. March (Ed.), *Handbook of organizations*. Chicago: Rand McNally, 1965.

MASLOW, A. H. *Motivation and personality*. New York: Harper, 1954.

MCGREGOR, D. *The human side of enterprise*. New York: McGraw-Hill, 1960.

PORTER, L. W. A study of perceived need satisfactions in bottom and middle management jobs. *Journal of Applied Psychology*, 1961, 45, 1–10.

PORTER, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *Journal of Applied Psychology*, 1962, 46, 375–384.

PORTER, L. W. *Organizational patterns of managerial job attitudes*. New York: American Foundation for Management Research, 1964.

PORTER, L. W., & LAWLER, E. E. Properties of organization structure in relation to job attitudes and job behavior. *Psychological Bulletin*, 1965, 64, 23–51.

(Received April 15, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 145–151

EFFECT OF SIMULATED SOCIAL FEEDBACK ON INDIVIDUAL TRACKING PERFORMANCE[1]

WILLIAM A. JOHNSTON AND LEON H. NAWROCKI

Ohio State University

Individual tracking performance was examined under conditions of simulated social feedback. Each of 60 Ss was told he had a partner and that posttrial feedback represented their team performance relative to average tracking ability. Actually, S's feedback represented his individual performance relative to a lenient, moderate, or stringent criterion. These criteria simulated partners of varying ability. Ss blamed their contrived partners for poor scores received under the stringent criterion. Performance of good trackers was not affected by criterion difficulty, but poor trackers performed best under the moderate criterion. The inhibitory influence of the stringent criterion was magnified during a terminal extinction session. The results suggest that criterion difficulty is an important determinant of performance in team and perhaps individual tasks.

Evaluative feedback that is some function of team output may be termed *social feedback*. Social feedback characterizes a wide array of interpersonal settings but has only recently received direct experimental attention (Burnstein & Wolff, 1964; Egerman, Glaser, & Klaus, 1963; Hall, 1957; Rosenberg, 1960; Shaw, 1962). An inherent aspect of social feedback is that it typically misrepresents individual performance. As an example, suppose that A and B are teammates, that A is a good performer and B a poor performer, and that they receive social feedback which represents their team performance relative to some criterion. Team performance is considered here to be some function of the average (possibly weighted) of the individual contributions. The criterion to which team performance is compared might be the performance of an average individual or team, the performance of a rival team, perfect performance, or any a priori standard established by the evaluator. Now, close scrutiny of this hypothetical team arrangement reveals that, in effect, A's good performance makes the criterion more lenient from B's point of view, but B's poor performance makes the criterion more stringent for A. That is, in comparison with evaluative feedback of individual performance, A has to perform better to obtain feedback of *good*, but B can perform worse. In short, social feedback underestimates A's individual performance, but overestimates B's performance.

The primary purpose of the present study was to determine if individual performance is affected by criterion difficulty. The major thesis is, of course, that social feedback can affect individual performance by controlling criterion difficulty. To the extent that the thesis is supported, individual performance under conditions of social feedback should be better understood. However, social feedback is not the only determinant of criterion difficulty: one evaluator may simply employ a more stringent a priori criterion than another; one competitor may be a better performer than another and thereby constitute a more stringent criterion; or perfect performance on one task may be more difficult to achieve than on another task. Furthermore, criterion difficulty is closely related to level of aspiration, which variable has received considerable experimental and theoretical interest (Locke, 1966; McClelland, Atkinson, Clark, & Lowell, 1953).

A subsidiary purpose of the present study was to ascertain any effect that social feed-

[1] The research was carried out at the Human Performance Center, and was sponsored by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant No. AF-AFOSR-985-66. Further reproduction is authorized to satisfy the need of the United States Government. The writers appreciate the assistance of Linda Lewis in collecting and analyzing the data.

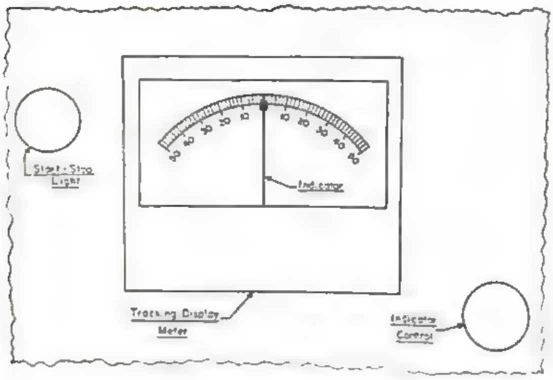


FIG. 1. Reproduction of a tracking display.

back might have on an individual's *evaluation* of his performance. Since individual contributions are difficult or impossible to disentangle from social feedback, A (of the hypothetical team) may form the impression that he performs worse than he really does, and B may be misled into believing that he performs better than he really does. Thus, social feedback may not only affect performance itself, but may affect self-evaluations of performance as well.

In this experiment, individual performance of a compensatory tracking task was examined under conditions of *simulated* social feedback. In actuality, each *S* performed alone and was *individually* evaluated relative to a criterion purported to represent average tracking ability. However, each subject (*S*) was told that he had a partner, and that his feedback represented the performance of his team relative to average tracking ability. Furthermore, criterion difficulty was varied experimentally at three levels (lenient, moderate, and stringent). Thus, it was *as if S* had a good partner under the lenient criterion, an average partner under the moderate criterion, and a poor partner under the stringent criterion.

METHOD

Subjects and Apparatus

The *Ss* were 60 male undergraduates, each of whom served five consecutive sessions and was paid $6.25 for his service. All *Ss* were naïve to experimental tracking tasks.

The apparatus was a specially constructed, compensatory tracking device which is described in detail in another report (Gain & Fitts, 1959). The tracking unit is diagrammed in Figure 1. The tracking display was a $2\frac{3}{8} \times 4\frac{1}{8}$-inch meter containing a zero-centered scale with a 50-unit range on either side of zero. An indicator could be set to sweep across the scale according to a predetermined wave pattern. The wave pattern was generated from two low-frequency (.1 cycle per second) sinusoids, and *S*'s task was to manipulate a control knob in an attempt to hold the indicator steady at zero.

The performance measure was the integrated absolute value of tracking error which was scored automatically over the final 55 seconds of each 60-second trial. The original scale was in terms of arbitrary voltage units. The score for each trial was later transformed from voltage units into a scale of linear extent (millimeter units). The transformed metric will henceforth be referred to as *average error* and represents the average deviation in millimeters which would be obtained if *S*'s error-amplitude distribution were drawn to the scale of the tracking display.

A feedback meter was mounted on the tracking unit and was identical to the tracking display except that the area to the left of zero was shaded red and designated *poor*, and the area to the right of zero was shaded green and designated *good*. The poor range was purported to represent below-average scores, and the good range to represent above-average tracking performance. Two complete tracking units were used and were housed in separate sound-treated booths (tracking-unit stations). The experimenter (*E*) controlled and scored the tracking task from still another room. Independent communication from *E* to each *S* was effected by two-way speaker systems in which channel direction was determined by *E*.

Design and Procedure

For experimental expediency, *Ss* were run two at a time, one *S* in each tracking-unit station. Each session was composed of twenty 60-second trials. The intertrial interval was 30 seconds, during which time *E* recorded *S*'s score and administered feedback. A green light on *S*'s unit automatically went on at the start of a trial and off at the trial's termination. A verbal ready signal was given immediately prior to the start of a trial.

Individual feedback of the raw error scores (in original voltage units) was verbally administered on the first two (practice) sessions. Sessions 3-5 constituted the experimental sessions. On Session 3 each *S* was assigned to one of three criterion-difficulty conditions: lenient, moderate, or stringent. Assignments were random with the restriction that there be an equal number of *Ss* per condition. All *Ss* received social feedback instructions on Session 3. There follows a descriptive portion of these instructions:

> Today you will become a member of a two-man team, and your scores will be averaged with

those earned by your teammate on corresponding trials. Actually, your partner has already completed the experiment and we are simply using the record of his scores. Nevertheless, your score averaged with his may be considered a true *team score* which reflects how well the two of you do *as a team*. After each trial your team score will be computed and then compared to a *norm*. The norm is simply the score earned by the average subject after the same amount of practice. Thus, you will be told how well your team performs relative to the average individual tracker. This comparison will be accomplished by means of the *feedback meter* mounted on your tracking unit. The zero point at the center of your feedback meter represents the norm, that is, average tracking ability. The green region to the right of zero indicates better-than-average scores, while the red region to the left of zero represents below-average performance. . . . Now, it should be made clear that your team score may misrepresent your own individual performance. Conceivably, you may be better than average as an individual but receive a team score in the red owing to a below-average partner; or you may be worse than average but receive a team score in the green because of an above-average partner.

Running mates were informed that they were *not* teammates. Social feedback was presented on *S*'s feedback meter for a 5-second duration approximately 15 seconds after each trial on Sessions 3 and 4. The three levels of criterion difficulty were determined from a distribution of error scores collected on Sessions 3–5 of a pilot study ($N = 16$). The first, second, and third quartiles of this distribution were selected as the stringent, moderate, and lenient criteria, respectively. These criteria were intended to simulate poor, average, and good partners. In terms of the arbitrary voltage scale employed, the criterion values were 35, 50, and 65, respectively, and the corresponding average error scores were 2.56, 3.65, and 4.75 millimeters. Social feedback was an expression of *S*'s error score relative to his criterion in original voltage units. In the event that *S*'s score was more than 50 units away from his criterion, his feedback needle would point to the appropriate extreme and *E* would issue the *off-meter* value verbally. Feedback was discontinued on Session 5 (extinction), but *S* was requested to estimate his own individual score (in feedback meter units) immediately after each trial.

RESULTS

An analysis of variance of Session 2 median average error scores revealed that neither tracking-unit stations, $F(1/54) < 1.00$, groups (which were not differentiated experimentally until Session 3), $F(2/54) < 1.00$, nor the interaction of these variables, $F(2/54) = 1.22$, were statistically significant ($p > .05$). These results justified the pooling of the data derived from the two tracking-unit stations for all subsequent analyses, and warrant the conclusion that the groups were equivalent prior to the introduction of experimental conditions.

In order to determine if effects of criterion difficulty on tracking performance depend on tracking ability, the 20 *S*s of each group were rank-ordered in terms of their median average error on Session 2. Then the eight best and eight worst trackers per group were designated *good* and *poor* trackers, respectively. Data of the middle four trackers of each group were omitted from subsequent analyses in order to maximize the separation of the two ability levels. This procedure yielded a 2 (Tracking Ability) × 3 (Criterion Difficulty) factorial design with eight *S*s per cell.

Evidence on the absolute difficulty of the three criteria is presented in Table 1. Due to an unfortunate lack of generality from pilot to experimental data, it is clear that all three criteria were more difficult than was intended. Thus, only the good trackers under the lenient condition tended to receive feed-

TABLE 1

MEDIAN FEEDBACK SCORES RECEIVED ON SESSIONS 3 AND 4 AS A FUNCTION OF ABILITY LEVEL AND CRITERION DIFFICULTY

Tracking ability	Session 3			Session 4		
	Lenient	Moderate	Stringent	Lenient	Moderate	Stringent
Good	+4.5	−18.5	−28.5	+18.5	−12.5	−14.0
Poor	−36.5	−42.0	−83.0	−23.5	−26.0	−62.5

Note.—Each score is the simple median of the pool of 160 scores derived from the eight *S*s (20 trials each) per condition. Positive and negative signs indicate scores above and below criterion, respectively.

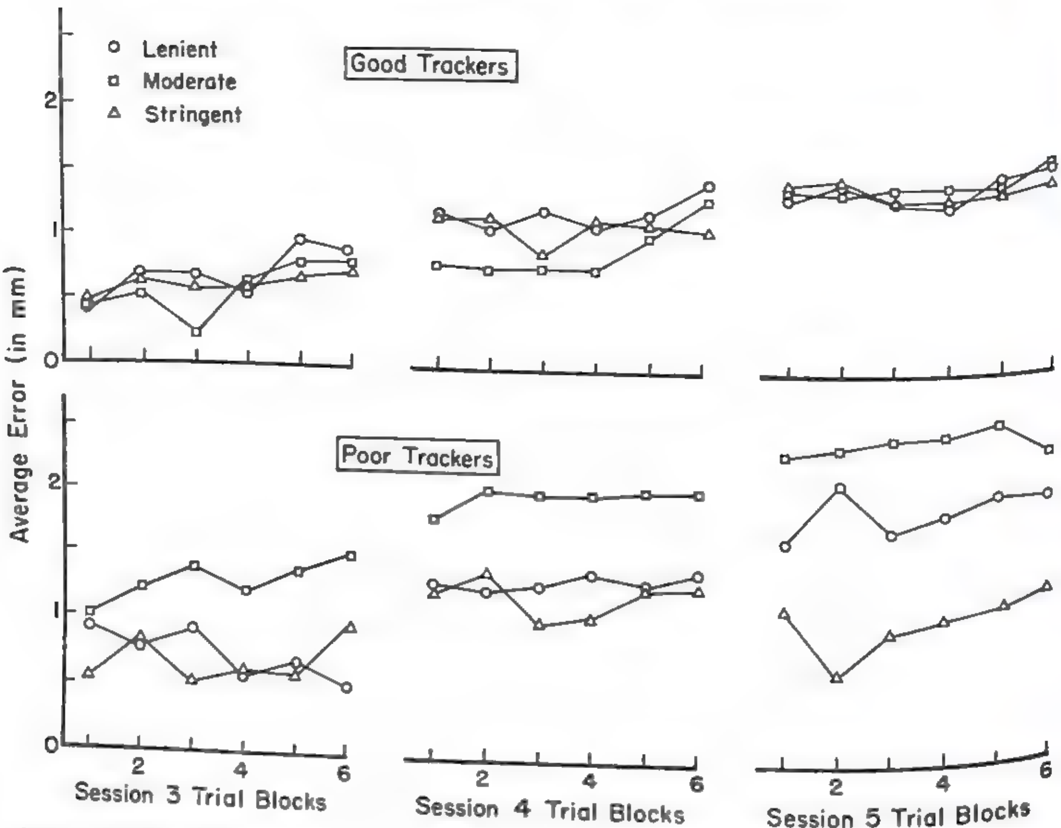


FIG. 2. Intra- and intersession improvement in average error as a function of tracking ability and criterion difficulty.

back scores in the *good* range, and, in fact, the poor trackers under the stringent criterion tended to receive all *off-meter* scores in the poor range. Thus, the three criteria were all stringent in an absolute sense, and were lenient, moderate, and stringent relative to one another only.

In order to eliminate "warm-up" effects, average error data for the first two trials of each session were omitted. The remaining 18 trials per session were organized into six three-trial blocks, and the simple median score of each trial block for each *S* was subtracted from his Session 2 median average error to yield an improvement score. This procedure eliminated the contaminating influence of extremely poor scores which occasionally arose when *S* explored a new tracking method or tested the scoring system by not tracking. Within- and between-session trends in amount of improvement as a function of ability level and criterion difficulty are portrayed in Figure 2. An analysis of variance of these data revealed that the amount of improvement was directly related to trial blocks, $F(5/210) = 6.71$, $p < .001$, and sessions, $F\ (2/84) = 69.33$, $p < .001$, and inversely related to tracking ability, $F\ (1/42) = 7.05$, $p < .05$. The first two were learning effects and the latter effect simply reflects that poor trackers had more *room* for improvement than had good trackers. Though criterion difficulty was not significant as a main effect ($.05 < p < .10$), it did interact significantly with tracking ability, $F\ (2/42) = 3.88$, $p < .05$. Duncan's test (Edwards, 1963) showed that the amount of improvement was equal under the three levels of criterion difficulty for the good trackers, but was greater under the moderate criterion than under the lenient and stringent criteria for the poor trackers ($p < .05$). The difference in the performance of poor trackers under moderate and lenient criteria is somewhat remarkable in view of the fact that these criteria differed only slightly in terms of ab-

TABLE 2

ABSOLUTE AND RELATIVE SELF-EVALUATIONS OF PERFORMANCE ON SESSION 5

Tracking ability	Absolute evaluations			Relative evaluations		
	Lenient	Moderate	Stringent	Lenient	Moderate	Stringent
Good	−3.42	+1.45	−6.30	−11.22	+1.33	+9.86
Poor	−13.82	−9.26	−21.79	−1.92	+6.04	+14.16

Note.—Each score is a mean evaluation score ($N = 8$). The derivation is explained in the text.

solute level of difficulty (see Table 1). Though the Sessions × Tracking Ability × Criterion Difficulty interaction only approached the .05 level of significance, F (4/84) = 2.17, $p < .10$, Duncan's test ($p < .01$) substantiated the important extinction effect indicated in Figure 2.[2] Specifically, differential effects of the lenient and stringent criteria on the performance of poor trackers emerged during extinction only. The differentiation of the three groups of poor trackers during extinction implies that criterion difficulty affects learning and not merely performance.

Mean absolute self-evaluations of Session 5 performance are presented in the left half of Table 2. An absolute self-evaluation score was simply the mean of S's estimates of his *individual* feedback scores. An analysis of variance of these data exposed tracking ability, F $(1/42) = 4.77$, $p < .05$, as the only significant source of variance. Thus, good trackers tended to give higher self-evaluations than poor trackers. However, it is noteworthy that even the self-evaluations of good trackers were low in an absolute sense. This may stem from the fact that their social feedback scores were also low (see Table 1), and is therefore clear evidence that social feedback distorted absolute self-evaluations. It is also noteworthy that the three criterion conditions yielded different feedback scores (see Table 1) but statistically equivalent self-evaluations, and that self-evaluations (especially of poor trackers) tended to surpass the corresponding social feedback scores.

These facts suggest that criterion difficulty affects S's appraisal of his partner's performance. Suppose, for example, that two Ss receive social feedback of −40 and −20, respectively, but that each estimates his individual score to be −20. It follows that the two Ss estimate their partners' scores to be −60 and −20, respectively. Thus, the difference between an S's social feedback score and his absolute self-evaluation score is an index of his *relative* (to partner) self-evaluation.

Mean relative self-evaluations of Session 5 performance appear in the right half of Table 2. Since social feedback was not given on Session 5, a relative self-evaluation score was derived by comparing S's mean absolute self-evaluation to the mean feedback score he would have received had social feedback been administered. A positive score indicates that S felt he did better relative to his criterion than his social feedback would have indicated, that is, that he felt superior to his partner. Conversely, a negative relative self-evaluation indicates that S felt inferior to his partner. Tracking ability, F $(1/42) = 10.52$, $p < .01$, criterion difficulty, $F(2/42) = 15.92$, $p < .001$, and the interaction thereof, F $(2/42) = 3.41$, $p < .05$, all attained significance in an analysis of variance performed on relative self-evaluation scores. As is apparent in Table 2, relative self-evaluations were higher for poor than for good trackers and were directly related to criterion difficulty. The interaction reflects a magnification of the tracking ability effect under the lenient criterion. These data indicate that S tends to accept good scores as accurate reflections of his own performance, but tends to attribute the blame for poor scores to his partner.

[2] The .01 level of significance was preferred to the more lenient .05 level in order to compensate for the fact that a nonsignificant interaction was being examined.

DISCUSSION

The data support the proposition that performance in a cooperative arrangement is affected by criterion difficulty. Individual performance of poor trackers was found to vary reliably with the ability level of the contrived partner, that is, with criterion difficulty. This effect is undoubtedly ubiquitous since social feedback, in obvious or subtle form, characterizes a wide spectrum of interpersonal settings ranging from the formal team to such loosely knit organizations as scientific or professional associations. Of course an invariant partner was simulated in the present study, and the observed performance effect of criterion difficulty may be altered if a more realistic partner is simulated by a variable level of criterion difficulty. Nonetheless, responses to a postexperimental questionnaire indicated that all but one of the *S*s believed that the feedback meter registered genuine team scores. Thus, the social instructions were sufficiently deceptive.

The data also suggest that one's evaluation of his contribution to team output is affected by the ability of his teammates. Specifically, *S* tended to accept the credit for good team scores but not the blame for poor team scores. Hence, criterion difficulty should be considered in any variable analysis of the formation of attitudes toward group members (Lott & Lott, 1960). Indeed, self-evaluations may play a mediating role in the performance effect of criterion difficulty. When team performance is good, each member shares the credit with his teammates, group morale is likely to be high, and individual effort should be maintained. However, when team performance is poor, each member attributes the blame to his partners, group morale is likely to be low, and individual effort may decline.

However, it is important to note that self-evaluation and average error scores did not completely covary in the present study. Thus, the effect of social feedback cannot be explained solely in terms of control of "social cognitions" by criterion difficulty. This implies that criterion difficulty may affect performance under asocial as well as social conditions. That is, criterion difficulty may influence individual performance even when *S* does not have a partner to whom he can attribute the blame for poor scores, or with whom he can share the credit for good scores.

In general, the present data imply that it is possible to maximize the performance (and possibly the morale) of a team by appropriate manipulation of its member composition. Furthermore, if criterion-difficulty effects are not restricted to social conditions, then criterion difficulty may influence behavior in any situation in which individual performance is evaluated relative to a criterion. The performance of an industrial or military worker may therefore be affected by the evaluative strictness of his supervisor, and a supervisor may be able to maximize the performance of his men by optimizing the stringency of his standards of evaluation.

Unfortunately, the optimal level of criterion difficulty cannot be specified until the more lenient end of the continuum of criterion difficulty is examined. However, the apparent inhibitory effect of the stringent criterion at least casts doubt on the notion that the higher one's goals (criteria) are, the better he performs (Locke, 1966). Furthermore, the data provide tentative information concerning the boundaries within which the effect of criterion difficulty is confined. Specifically, it appears that the effect is on learning as well as performance and is most likely to occur with relatively unskilled performers. Finally, the sheer demonstration that a criterion-difficulty effect exists is important in view of the fact that the informational value of feedback is independent of criterion difficulty. Hence, the present findings contribute to a data base from which an explanation and full appraisal of criterion-difficulty effects may eventually be developed.

REFERENCES

BURNSTEIN, D. D., & WOLFF, P. C. Shaping of three-man teams on a multiple DRL-DRH schedule using collective reinforcement. *Journal of Experimental Analysis of Behavior*, 1964, 7, 191-199.

EDWARDS, A. L. *Experimental design in psychological research.* New York: Holt, Rinehart & Winston, 1963.

Egerman, K., Glaser, R., & Klaus, D. J. Increasing team proficiency through training: 4. A learning-theoretic analysis of the effects of team arrangement on team performance. *American Institutes for Research Report*, 1963, No. B64-9/63-TR.

Gain, P., & Fitts, P. M. A simplified electronic tracking apparatus (SETA). *USAF WADC Technical Report*, 1959, No. 59-44.

Hall, R. L. Group performance under feedback that confounds responses of group members. *Sociometry*, 1957, 20, 297–305.

Locke, E. A. The relationship of intentions to levels of performance. *Journal of Applied Psychology*, 1966, 50, 60–66.

Lott, B. E., & Lott, A. J. The formation of positive attitudes toward group members. *Journal of Abnormal and Social Psychology*, 1960, 61, 297–300.

McClelland, W. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.

Rosenberg, S. Cooperative behavior in dyads as a function of reinforcement parameters. *Journal of Abnormal and Social Psychology*, 1960, 60, 318–333.

Shaw, M. E. Implicit conversion of fate control in dyadic interaction. *Psychological Reports*, 1962, 10, 75–78.

(Received April 27, 1966)

Manuscripts Accepted for Publication in the *Journal of Applied Psychology*

Development of a Method of Predicting High Accident and High Violation Drivers: Bruce D. Greenshields* and Fletcher N. Platt: Department of Civil Engineering, 1201 E. Engineering Building, University of Michigan, Ann Arbor, Michigan.

Determinants of Satisfaction in Middle-Management Personnel: Dilip K. Lahiri and Suresh Srivastva*: Alfred P. Sloan School of Management, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, Massachusetts 02139.

Goal Setting as a Means of Increasing Motivation: Judith F. Bryan and Edwin A. Locke*: American Institutes for Research, 8555 Sixteenth Street, Silver Spring, Maryland 20910.

The Graphoanalytic Approach to Selecting Life Insurance Salesmen: S. M. Zdep* and H. B. Weaver: Department of Psychology, University of Hawaii, Honolulu, Hawaii 97 96822.

Worker Preferences among Time-Off Benefits and Pay: Stanley M. Nealey* and James G. Goodale: Department of Psychology, University of Illinois, 301 Gregory Hall, Urbana, Illinois 61801.

Measurement of Industrial Relations Activities: Philip Ash*: Inland Steel Company, 30 West Monroe Street, Chicago, Illinois 60603.

(Continued on page 180)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 152-160

DIFFERENCES BETWEEN IDENTIFIED AND ANONYMOUS SUBJECTS IN RESPONDING TO AN INDUSTRIAL OPINION SURVEY

STUART M. KLEIN
IBM Corporation

JOHN R. MAHER
Columbia University

AND RICHARD A. DUNNINGTON
IBM Corporation

A comparison of attitude survey responses between identified and nonidentified manufacturing employees was made under 2 conditions of identification. 1 involved a face-to-face designation by the respondee's manager as to which group he was to be in (high threat), and the other involved a random allocation as the respondee entered the testing room (low threat). All Ss were assured confidentiality of their responses, and the nonidentified respondees were assured anonymity. A positive distortion in responses took place under both identified conditions, but significantly more under high threat. Moreover, the items themselves produced variable distortion. Items dealing with salary and with ratings of top management produced consistent positive distortions, whereas items dealing with work pressure and S's manager produced little or no distortion even under conditions of high threat.

One of the major problems faced by behavioral scientists whose principal research method involves attitude survey work is to determine cause and effect. Because the "causal" variables are measured in the same instrument as the dependent variables the two classes are obviously not independent from each other and are consequently susceptible to a systematic bias. Moreover, relationships that occur among the variables may be due to causal factors or to conceptual similarity, and, while assumptions can be made as to which is which, these assumptions can rarely be pinned down with a high degree of certainty. Finally, there is always the dilemma of determining antecedents and consequences.

These problems have received the attention of a number of methodologists who have suggested ways of dealing with them (Blalock, 1961; Hyman, 1955). One of the ways that has been used is to identify the subjects (*S*s) and then tie up certain demographic variables to attitudinal responses both at one point in time and longitudinally. In this way independent and antecedent variables are identified and a limited spectrum of causality can be determined. However, this method introduces a new variable; namely the fact of having been identified. The effects of identification on responses to questionnaire items constitute the major focus of this paper.

Specifically, the issues that are dealt with include: (*a*) whether *S*s will indeed respond differently when their names are on the answer sheet than when they are anonymous; (*b*) if so, whether different conditions of identification can lead to different degrees of "faking"; and (*c*) whether *S*s who do have a tendency to try and supply "right" answers will do so on all items or just certain ones.

Much of the experimental work that has been done on "faking" concerns personality inventories, such as the MMPI, and deals directly only with Problem (*a*)—that is, can and do *S*s "fake" successfully in paper-and-pencil tests? Within this body of research there is a lack of complete unanimity in answering the question. Sheldon (1959) points out that studies concerned with whether or not an individual can successfully distort his score favorably have been confounded by use of different items and scales, noncomparable *S* populations, and different experimental conditions. Moreover, Callis (1950) and Rabinowitz (1954), using exactly the same inventory and comparable *S*s, obtained diametrically opposed results in attempting to answer the question of whether "faking"

was possible on the Minnesota Teacher Attitude Inventory.

A number of studies report an apparent absence of response distortion by identified *Ss* (e.g., Ash & Abramson, 1952; Corey, 1937; Gerberich & Mason, 1948; Hamel & Reif, 1952; Olson, 1936). All but the Hamel and Reif study used college students as *Ss*. Pelz (1959) compared the effects of full anonymity versus confidentiality, with the latter condition being defined by individual identification to the survey staff *only*. Although practically no differences obtained between the two conditions, the author qualifies his findings because of the special nature of the organization tested. According to Pelz,

> (1) The personnel studied were largely covered by civil service regulations, which provide high job security.
>
> (2) In general, the personnel had high confidence in management's integrity and concern for employee welfare.
>
> (3) In general, the personnel had confidence in the integrity of the outside survey staff and in the confidentiality of the data. This was built up over several months of spadework [p. 90].

Rosen (1960) investigated the effects of identification on attitudes expressed by college students toward (*a*) reading in general and (*b*) a specific English course given in the university they attended. Although he found statistically significant differences between identified and anonymous groups in attitudes expressed toward the specific English course, the difference between the two group means was only 1.52 on a 40-point scale. There were no significant differences in attitudes toward reading in general. Thus Rosen concludes

> On balance, the literature and the study reported here strongly suggest that identification of respondents in attitude questionnaire surveys conducted under less than highly threatening circumstances is not likely to result in serious statistical or practical distortion. . . . Even in the few studies where statistically significant differences between anonymous and identified respondents have been reported, the magnitude of those differences has been inconsequential for practical purposes [p. 679].

Thus there is some indication that there is no appreciable difference between the responses of identified and nonidentified *Ss*. However, there is also evidence that suggests a difference exists. For example, Green (1951) states:

> The literature that is relevant to the problem is very meager. But what does exist indicates quite definitely that overzealous applicants for specific positions in real life *can*, *will*, and *do* bias their answers on tests that are amenable to such behavior—especially if they think they will get away with it [p. 505].

In addition, Wesman (1952) found that the same group of 73 *Ss* responded significantly differently when assuming two different response sets at different times in answering the same questionnaire. Furthermore, these two sets of answers were judged to have a high degree of face validity when matched against the two profiles *Ss* had been asked to assume.

Kahn (1951) and Metzner and Mann (1952) obtained significant differences between attitudes expressed in an anonymous questionnaire and in a face-to-face interview, while Elinson and Haines (1950) and Benson (1941) report similar findings in comparing anonymous and identified responses to the same questionnaire. However, since Elinson and Haines were dealing with a population of army enlisted men, while Benson surveyed residents of a small Republican community in Maine about voting preferences, it is possible that the identified *Ss* in these latter two studies were confronted with implicit pressures to respond the "right" way.

Dunnette and Heneman (1956) dealt more directly with the notion of perceived threat as a determinant of whether identified *Ss* will distort their responses favorably. In their experiment, all responses were literally anonymous, but "psychological anonymity" was varied. In one condition, an attitude questionnaire was administered to *Ss* by a research staff member from outside *Ss*' organization, while in a second condition it was administered by an official of the firm in which the survey was being conducted. It was found that (*a*) where employees feelings of anonymity were threatened by the presence of an official of the company, responses to the questionnaire were more favorable; (*b*) differential amounts of response distortion were obtained depending upon the content of the items; and (*c*) employees experiencing a

TABLE 1

Comparison of Identified and Anonymous Samples in Plant G on Demographic Items

Item	Identified			Anonymous			
	X	σX	N	X	σX	N	t
Length of service	4.08	1.46	410	4.03	1.41	4.28	0.50
Education	3.21	0.69	404	3.20	0.69	4.27	0.21
Shift	1.50	0.63	411	1.49	0.61	4.28	0.23
Age	2.33	1.05	411	2.27	1.08	4.29	0.82
Marital status	1.20	0.40	411	1.18	0.38	4.28	0.74

threat to anonymity gave fewer and shorter write-in comments.

If one postulates that threat is the underlying cause of the degree of distortion obtained in some of the referenced studies, one would also expect that the manner in which Ss are identified would have an effect. If Ss are identified in a totally impersonal manner, we would expect less faking than if they are told face-to-face that their questions would have an identifying characteristic even though under both conditions of identification Ss are assured of the confidentiality of their responses.

Finally, in accordance with Dunnette and Heneman (1956), not all items in an attitude survey should be of equal threat. One might expect that items dealing with things of a more personal factor such as attitudes about an individual's manager would be more threatening than those dealing with impersonal factors such as working conditions.

In summation, three questions have been addressed:

1. Do Ss who are identified respond differently to a company-sponsored attitude questionnaire than Ss who remain anonymous?

2. Does the manner of identification have an effect on the degree of difference?

3. Does the kind of item that S is responding to affect the degree of this difference?

Method

Subjects

The Ss included all manufacturing nonmanagement employees from Plant D and Plant G, two manufacturing installations in a large industrial corporation.

The Questionnaire

A 290-item questionnaire was administered to Ss in both locations. This questionnaire was a general attitude inventory centered around Ss' occupation, including such areas as attitudes toward salary, job, and management.

Independent Variables

Identified versus anonymous subjects. In both plants approximately one-half of the Ss were told that they should write down a personal identifying number. The other Ss were told to check a departmental code number but not to write down a personal identifying number. In all cases Ss were assured that no attempt would be made to identify anyone to management. Only people in Personnel Research would ever make use of the personal identifications and then only for research purposes. In other words, confidentiality was assured, though anonymity in the case of identified Ss was not.

High versus low threat in the identification procedure. The high-threat manipulation took place in Plant D. Under the high-threat conditions the identified Ss were informed in their department by their manager that they would have to put down an identifying number on their questionnaires when they received their booklets in the auditorium. There were 111 employees in both the experimental and control groups.

The low-threat manipulations took place in Plant G. Under the low-threat condition, the manager said nothing to the identified Ss. The process of identification occurred by assigning every other S to the identified group as they walked into the room where the questionnaire was administered. There were 411 employees in the experimental group and 430 in the control group.

In both plants, the completed questionnaires were collected in an identical manner. Both identified and anonymous Ss were told that when they had finished answering the questions, they were to place the questionnaire booklet in a cardboard box at the front of the room. Identified and anonymous questionnaires were collected in the same box, and the administrator was careful to avoid giving any respondent the impression that he was "checking up" on any of the questionnaires at any time during administration or collection. In addition, the method of identification insured that no S's name would appear on his booklet. After an S had deposited his answers in the collection box, he was free to leave the room and return to work.

In Plant D, the identified and anonymous groups were matched prior to the test for demographic characteristics such as age, education, and length of service. In Plant G, where random assignment was a necessary part of the low-threat manipulation, such information was collected on the questionnaire and the two groups were compared a posteriori for significant differences. Table 1 shows that the two

groups in Plant G were, for all practical purposes, equivalent and that random selection had truly taken place.

Data Analysis

Within both plants, the data were broken down for each item according to whether the respondent was identified or anonymous. Statistical significance between the groups was originally determined using a t test on the ensuing means for each of the items.

For several reasons, a substantial number of items had to be excluded a priori from the data analysis. Of the original 290 items, 114 were retained which met the following criteria: (*a*) The items dealt with subjective, attitudinal areas. One hundred thirty items had to be discarded because they were basically factual or dealt with objective reality, for example, all demographic items. (Subsequent analysis revealed little or no differences between groups in either threat condition on these items. A 10% representative sample of these discarded items appears in Table 2.) (*b*) The items included the same possible response categories in both locations. Twenty-two items did not meet this criterion. (*c*) The items employed linear scales which were directional with regard to favorability, thus making a comparison of mean differences a meaningful one. Three items did not have linear scales. (*d*) The items were responded to by at least half the *Ss* in each of the four conditions. Twenty-one items (dealing with change over time) could not be answered by at least half the *Ss* in each condition and had to be excluded from the analysis.

RESULTS

It is clear that identifying *Ss* makes a difference in their responses to the questionnaire. Under both conditions of threat a substantial number of items were responded to significantly more favorably by the identified group. Using a one-tailed t test with a significance level of .05 as the criterion of significance, one finds that 39% of the items are responded to more favorably under the

TABLE 2

TEN PERCENT REPRESENTATIVE SAMPLE OF ITEMS DISCARDED DUE TO BEING PRIMARILY OBJECTIVE

Item	Plant D (high threat)							Plant G (low threat)						
	Identified			Anonymous				Identified			Anonymous			
	X	σX	N	X	σX	N	t	X	σX	N	X	σX	N	t
Once you have fully mastered a particular job, what would you prefer to do? (Stay with the same job or transfer?)	1.90	0.30	111	1.87	0.33	111	0.73	1.59	0.49	409	1.54	0.50	428	1.46
How important does management consider doing work of high quality in deciding salary increases for employees in your department?	1.36	0.75	111	1.32	0.62	111	0.43	1.45	0.80	410	1.46	0.76	427	0.19
How many people in your department talk to each other when work permits?	1.63	0.81	111	1.60	0.61	111	0.31	1.72	0.80	410	1.76	0.82	428	0.71
Can the kind of work you usually do be accurately measured or counted?	1.25	0.44	110	1.36	0.48	111	1.78*	1.44	0.50	410	1.43	0.50	426	0.29
How many different jobs in this division have you worked at enough to be fairly competent?	2.95	1.37	111	2.68	1.40	111	2.16**	2.63	1.30	411	2.61	1.29	429	0.22
How much information do you usually have about how much work you are supposed to do in a given amount of time?	1.92	0.53	109	1.94	0.64	110	0.25	2.15	0.57	394	2.17	0.59	412	0.49

TABLE 2—(*Continued*)

Item	Plant D (high threat)							Plant G (low threat)						
	Identified			Anonymous				Identified			Anonymous			
	X	σX	N	X	σX	N	t	X	σX	N	X	σX	N	t
I would most prefer (1) a job where my instructions are quite detailed and specific or (2) a job where my instructions are very general.	1.40	0.49	111	1.43	0.50	111	0.45	1.46	0.50	406	1.47	0.50	424	0.29
How frequently does your manager come around to see how things are going?	3.20	0.59	110	3.11	0.77	111	0.98	3.23	0.62	410	3.23	0.74	427	0.00
How many suggestions have you submitted to the suggestion program during the past year?	3.61	1.34	110	3.46	1.37	110	0.82	4.00	1.28	407	3.96	1.19	424	0.47
Think back to the last time you considered seeking a transfer to another job. (To what) degree (did an opportunity to improve your skills lead) you to consider such a change?	1.52	0.64	104	1.45	0.57	108	0.84	1.65	0.64	355	1.67	0.66	375	0.42
How much "say" do you think (higher plant management) has in deciding the amount of work you should do on your job?	2.98	1.41	110	2.92	1.34	111	0.32	2.33	1.27	395	2.30	1.28	416	0.33
How important would interesting and challenging work be to you in choosing a job?	3.38	1.36	104	3.41	1.33	102	0.16	3.54	1.33	393	3.65	1.37	397	1.15
How important are company medical benefits to you?	2.38	1.31	107	2.57	1.19	107	1.11	2.53	1.13	400	2.52	1.14	416	1.26

* $p < .05$.
** $p < .025$.

condition of high threat and 55% of the items are responded to more favorably by those identified under the low-threat condition. Under both conditions the chi-square values are significant beyond the .001 level (Table 3).

Another way of looking at the data involves comparing the identified and anonymous groups in terms of the percentage of *Ss* responding in the two most favorable and/or two least favorable response categories on each item. Table 4 shows the mean percentage differences in these categories for both conditions of threat, as well as the range of the differences. Again, the effect of identification is clearly in evidence.

While these results demonstrate the presure of a strong effect associated with identification of respondents, they are also somewhat contradictory to the expectation that personal identification (high threat) will produce a greater effect than impersonal (low threat) identification. Cursory examination of the results indicates that the opposite happens. In the low-threat condition, 55% of the items discriminate between identified and

anonymous *Ss* and only 39% of the high-threat condition do so.

Further analysis of the data, however, shows this initial impression to be in error. The above differences are attributable to the fact of having a vastly different-sized sample in Plant D (high threat) than in Plant G (low threat). Since a total sample of 222 *Ss* was used in Plant D as opposed to 841 in Plant G, application of the standard t test requires a much greater mean difference between experimental and control groups in Plant D than in Plant G to achieve a given level of statistical significance.

Fortunately, there is a statistic which provides us with an estimate of the percentage of the total variance accounted for by a given effect. We refer to Hays' (1963) omega-square, given by the author as follows:

$$\text{omega-square} = \frac{t^2 - 1}{t^2 + N_1 + N_2 - 1}$$

Let us consider only those items where, in at least one of the populations, the variance contributed by the effect of the manipulation is greater than the error variance; in other words, where t is greater than 1 for one or both of the plants. There are, in all, 99 such items. According to the null hypothesis, one should expect (given that the omega-squares for Plant D and Plant G are never equal) 50 cases in which the omega-square in Plant D is larger than the omega-square in Plant G, and 49 cases where the reverse is true. However, one finds that there are actually

TABLE 3

NUMBER OF ITEMS DISCRIMINATING BETWEEN IDENTIFIED AND ANONYMOUS SUBJECTS AT THE .05 LEVEL

Condition	Discriminating items	Nondiscriminating items
High threat (Plant D)[a]		
Expected by chance	6	108
Observed	44	70
Low threat (Plant G)[b]		
Expected by chance	6	108
Observed	63	51

[a] $\chi^2 = 253.7$, $p < .001$.
[b] $\chi^2 = 571.58$, $p < .001$.

TABLE 4

DIFFERENCES IN PERCENT DISTRIBUTIONS OF IDENTIFIED VERSUS ANONYMOUS SUBJECTS

Condition	Two most favorable response categories	Two least favorable response categories
Plant D (high threat)		
M percent difference	+6.65	−5.09
Range of percent differences	−71 to +21.6	+4.1 to −18.2
Plant G (low threat)		
M percent difference	+4.11	−2.64
Range of percent differences	−39 to +12.2	+2.5 to −9.0

Note.—Differences = Identified on two categories minus Anonymous in these two categories.

61 cases of the former type and only 38 of the latter. This difference yields a chi-square of 5.16, significant at the .025 level with 1 degree of freedom. In fact, then, the high-threat manipulation yields a significantly greater effect than the low-threat manipulation.[1]

The remaining question to be answered is, does the nature of the item itself have anything to do with determining response distortion? In an attempt to resolve this issue, 60 representative items for the Plant G population were factor analyzed.[2] These items were selected a priori from three general types of items: (*a*) those which on the surface did not appear to be threatening but which discriminated between identified and anonymous *Ss*; (*b*) those which appeared threatening and did not discriminate; and (*c*) those which appeared to be threatening and did indeed discriminate.

Six significant factors emerged. They follow along with representative items in each:

[1] The Plant G population was selected rather than the Plant D population since the effects of the severe high-threat manipulation were originally expected to generalize to items which would not ordinarily be threatening enough to yield a significant difference.

[2] In the high-threat condition, values for ω^2 for the items in question ranged from .00095 to .10390. Under low threat, the range was from .00020 to .01689.

TABLE 5

PERCENTAGE OF ITEMS LOADING AT LEAST .30 ON EACH FACTOR THAT DISCRIMINATE BETWEEN IDENTIFIED AND ANONYMOUS SUBJECTS UNDER BOTH CONDITIONS OF THREAT

Factor	No. items loading .30 or higher	Percentage of items that discriminate
Attitudes toward company and top plant management	13	54
Ratings of immediate manager	8	0
Extrinsic job factors and grievance channels	3[a]	67
Attitudes toward salary and advancement	6	67
Intrinsic job factors	6	33
Workload	3	0

[a] 7 additional items with a loading of at least .30 on this factor in Plant G did not meet the criterion of 50% response in Plant D.

Factor I. *Attitudes toward the company and top plant management*. (19 items had primary loadings on this factor.)

1. How would you rate top plant management on fairness in dealing with employees?

2. How would you rate the plant as a place to work?

3. How would you rate top plant management on recognizing superior performance?

Factor II. *Ratings of immediate manager*. (8 items had primary loadings on this factor.)

1. How would you rate your manager?

2. How would you rate your manager on assigning the right jobs to the right people?

Factor III. *Extrinsic job factors and grievance channels*. (11 items had primary loadings on this factor.)

1. How often is management fair in listening to employee complaints?

2. Rate your job on how well work is scheduled.

Factor IV. *Attitudes toward salary and advancement*. (9 items had primary loadings on this factor.)

1. Rate your pay considering duties and responsibilities.

2. How satisfied are you with your advancement so far in the company?

Factor V. *Intrinsic job factors*. (9 items had primary loadings on this factor.)

1. How satisfied are you that your ability and training is properly used?

2. How do you like your job—the kind of work you do?

Factor VI. *Workload*. (5 items had primary loadings on this factor.)[3]

1. How do you feel about the amount of work expected of employees?

2. Suppose you did more work than required for the next few months . . . would your manager expect you to do still more?

Of the six factors three appear high in threat and two low in threat.[4] Using .30 primary loading as an arbitrary criterion, one finds that of all the items that meet this criterion for Factor IV (salary) 67% discriminate between the identified and nonidentified groups (Table 5). It is inferred, therefore, that items dealing with salary present a considerable threat to people whose responses can be identified. This inference is supported by the fact that two of the items with primary loadings of .30 or better on Factor I that discriminate under low threat are either directly or tangentially related to money matters (i.e., "How good a job does management do in evaluating employees for merit pay increases?" "How would you rate the plant and its top manager on providing a chance for employees to get ahead?").

Factor I (rating top plant management) produces seven items that discriminate and six which do not. Two of the items that are apparently unthreatening are "How would you rate top plant management on: (a) 'Fairness in dealing with employees?' and (b) 'Consideration of employees as individuals?'" On the other hand, items with close conceptual similarity such as, "How would you rate top plant management on the way employee complaints are handled?" do discriminate and thus seem to be relatively threatening.

Labeling Factor III (extrinsic job factors and grievance channels) as "highly threatening" seems justified but somewhat tenuous.

[3] The total of primary loadings adds up to 61 because 1 of the items had a split primary loading.

[4] Only those items that discriminate under both conditions of threat will be dealt with.

Of the original items which loaded at least .30 on this factor in the Plant G factor analysis, all but three had to be discarded because they were not responded to by 50% of the two subsamples in Plant D. The discarded items dealt with change over time, and somewhat less than 50% of the Plant D population had been employed long enough to answer them meaningfully and were told to skip them. Thus one hesitates to designate Factor III as being of big threat, even though two of the three items loading on it do discriminate between the identified and anonymous subsamples in both plants.

The least threatening conceptual areas are those tapped by Factor II (ratings of immediate manager) and Factor IV (workload and work pressure). None of these items with primary loadings of .30 or better on these factors discriminate between the identified and nonidentified groups even under the high-threat conditions. Factor V (intrinsic job factors) seems only moderately threatening at most. Only two of the six items are threatening enough to discriminate between identified and anonymous *Ss* in both conditions of threat.

Discussion

From a methodological viewpoint, the findings of this study are reasonably clear-cut. The data indicate that if we wish to identify nonmanagement manufacturing people in a company-sponsored attitude survey, we may be safe in doing so only up to a point. While there is no question that identification is threatening to the employees and thus leads to response distortion, the degree of distortion depends upon the way identification takes place and the kinds of questions asked. When the individual is personally told that his questionnaire must bear an identifying mark, then a considerable amount of "faking" takes place even though he is assured confidentiality for his responses. On the other hand, when he is randomly assigned to an identified or nonidentified group as he enters the examination room, this may be sufficiently less threatening so as to produce comparatively less "faking." Apparently the personal contact lowers the individual's anxiety threshold to a point where even marginally threatening items sometimes arouse sufficient feelings of threat to induce distortion in the socially desirable direction. When this personal acknowledgment of identification is absent the fact of identification usually does not, in and of itself, lower the anxiety threshold enough to create response distortion, except where the items themselves are intrinsically threatening.

Less clear is the degree of threat that seems inherent in different conceptual types of items analyzed. In accordance with Dunnette and Heneman (1956) we find that some items produce more distortion than others. We find, for example, that items dealing with pay and with the adequacy of evaluations made by management with regard to the individual's skills and abilities are sufficiently threatening to produce distortion under even the low-threat condition of identification. On the other hand, items dealing with work pressure and contrary to the Dunnette and Heneman findings with attitudes toward the individual's manager do not produce much distortion at all. A simple answer as to why differential distortion occurs might be that there are certain things on the job that are "OK" topics for complaint. For example, as long as the employee is giving 100% in terms of effort and productivity, who could object if he merely observed that he had more work to do than perhaps was reasonable? The worker has no control over the amount of work that is assigned to him, and consequently takes no personal responsibility for it; as long as he imagines himself as giving maximum effort, he does not fear negative sanctions if he complains about the amount of work inherent in his job.

Certain other areas, however, may have acquired the mystique of being taboo. Salary is a prime illustration of this. For example, *Ss* in interviews that we have conducted will rarely supply a direct answer to the question, "To whom do you compare yourself when evaluating your salary?" Instead, they will deny comparing themselves with anyone, even though later on in the interview they may prefix a comment about their pay with "Well, compared to. . . ." Perhaps employees perceive top management as more sensitive to criticism concerning their salary programs in general.

Moreover, in our society it is probably less legitimate to complain openly about money matters or the way one is evaluated. Both involve opinions about one's personal worth and require very subjective arguments in support of these opinions. These arguments are therefore difficult to substantiate and in the absence of substantiation could render the individual open to direct attacks upon his sense of personal worth. If one felt that there was any possibility of an open confrontation of this sort he might be constrained to distort his responses in order to avoid such a threat.

A major anomaly of these results is the lack of positive distortion among the identified *S*s in rating their manager. Certainly if our reasoning that threat is an important factor in determining the degree of distortion were correct, one would expect ratings of one's manager to be highly threatening and therefore conducive to more distortions as Dunnette and Heneman (1956) find. Perhaps the difference in the two findings is due to the differences in the manipulations performed. Whereas Dunnette and Heneman have a condition of "psychological" anonymity, we do not, since the questionnaires in our study were all administered by someone in the company. A positive bias might have already taken place with regard to ratings of immediate management in our population. Consequently, there may have been less room for an additional biasing effect.

In any event, a full treatment of this subject is beyond the scope of this research. Suffice to conclude that questionnaire items do contain varying degrees of threat and consequently are open to varying degrees of distortion. Thus, when attitude survey work is conducted, be it with questionnaires or interviews, care must be taken in both content and method to assure minimum distortion due to threat associated with the identification of the respondees. If the conditions of personal threat are low, then somewhat less distortion may be expected. But even so, items dealing with salary and ratings of top plant management may produce significant positive distortion.

REFERENCES

Ash, P., & Abramson, E. The effect of anonymity on attitude-questionnaire research. *Journal of Abnormal and Social Psychology*, 1952, **46**, 722–723.

Benson, L. E. Studies in secret ballot technique. *Public Opinion Quarterly*, 1941, **5**, 79–82.

Blalock, H. M., Jr. *Causal inferences in nonexperimental research.* Chapel Hill: University of North Carolina Press, 1961.

Callis, R. Change in teacher-pupil attitudes related to training and experience. *Educational and Psychological Measurement*, 1950, **10**, 718–727.

Corey, S. M. Professed attitudes and actual behavior. *Journal of Educational Psychology*, 1937, **28**, 271–280.

Dunnette, M. D., & Heneman, H. G., Jr. Influence of scale administrator on employee attitude responses. *Journal of Applied Psychology*, 1956, **40**, 73–77.

Elinson, J., & Haines, V. T. Role of anonymity in attitude surveys. *American Psychologist*, 1950, **5**, 315.

Gerberich, J. B., & Mason, J. M. Signed versus unsigned questionnaires. *Journal of Educational Research*, 1948, **42**, 122–126.

Green, R. F. Does a selection situation induce testees to bias their answers on interest and temperament tests? *Educational and Psychological Measurement*, 1951, **11**, 503–516.

Hamel, L., & Reif, H. G. Should attitude questionnaires be signed? *Personnel Psychology*, 1952, **5**, 87–91.

Hays, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

Hyman, H. *Survey design and analysis.* Glencoe, Illinois: The Free Press, 1955.

Kahn, R. A comparison of two methods of collecting data for social research: The fixed alternative questionnaire and the open-ended interview. Unpublished doctoral dissertation, University of Michigan, 1951.

Metzner, H., & Mann, F. C. A limited comparison of two methods of data collection: The fixed alternative questionnaire and the open-ended interview. *American Sociological Review*, 1952, **17**, 486–491.

Pelz, D. C. The influence of anonymity on expressed attitudes. *Human Organization*, 1959, **18**, 88–91.

Olson, W. C. The waiver of signature in personal reports. *Journal of Applied Psychology*, 1936, **20**, 442–449.

Rabinowitz, W. The fakability of the Minnesota Teacher Attitude Inventory. *Educational and Psychological Measurement*, 1954, **14**, 657–664.

Rosen, N. A. Anonymity and attitude measurement. *Public Opinion Quarterly*, 1960, **24**, 675–679.

Sheldon, M. S. Conditions affecting the fakability of teacher-selection inventories. *Educational and Psychological Measurement*, 1959, **19**, 207–210.

Wesman, A. G. Faking personality test scores in a simulated employment situation. *Journal of Applied Psychology*, 1952, **36**, 112–113.

(Received April 28, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 161–166

VISION AND KINESTHESIS IN THE ACQUISITION OF TYPEWRITING SKILL [1]

LEONARD J. WEST

City University of New York (Division of Teacher Education)

266 Ss at typing skill levels from 9 through 108 wpm typed from ordinary prose under visual and nonvisual conditions, under instructions to indicate when they sensed having made an error. Results confirmed the hypothesis of a swift rise in kinesthetic dependability from low levels among beginning typists but showed a plateau from intermediate through expert levels of skill; an r = .36 was found between kinesthetic dependability and skill level. Kinesthetic feedback was at significantly lower levels than all-senses feedback throughout the range of typing skill, while deprivation of vision had no effects on speed but resulted in large and significant increases in errors. These findings suggest the free use of vision in early stages of learning to typewrite, as contrasted with the conventional insistence on so-called "touch" operation at the start.

Concerning the sensory processes that underlie the acquisition of complex perceptual motor skills, Honzik (1932) and Fitts (1951) had hypothesized that kinesthetic cues come into play as a basis for making and for confirming responses only after some learning has taken place on the basis of exteroceptive stimuli, primarily vision. Taking the typewriting task as a prominent example of a skill presumed to depend increasingly on utilization of kinesthetic cues, this investigation had as one of its two major purposes estimation of the level of dependable kinesthetic feedback at successive stages of typing skill and, in turn, of the relationship between level of skill and extent of utilization of kinesthetic cues. What is the extent of the hypothesized shift from dependence on exteroceptive toward dependence on proprioceptive stimuli as skill is acquired? Fleishman and Rich (1963), in their investigation of spatial and kinesthetic sensitivity, showed superior performance later in learning a two-hand coordination task by those high in kinesthetic sensitivity, thus furnishing tangential support for Fitts' hypothesis, but not any direct measures of extent of dependability of kinesthetic cues at successive stages of skill.

The second major purpose of this investigation was to assess the effects on performance (speed and errors) of deprivation of vision. Earlier studies (e.g., Carr, 1927; Koch & Ufkess, 1926; Zigler, 1932) had routinely shown deprivation of vision to have large negative effects on acquisition in maze learning. Insofar as the conventional instructional insistence on so-called "touch" typing from the start of learning flatly contradicts the existing hypothesis and the accumulation of evidence for other perceptual motor tasks, it was thought desirable to collect direct evidence for the typing task.

Specifically, then, the primary objective was to test the hypothesis of increasing dependability of kinesthetic cues as skill is acquired and, in turn, to estimate the relationship between typing skill level and extent of dependable kinesthetic feedback. That is, if, as performance becomes habitual, proprioceptive feedback or "feel" comes into increasing play, then kinesthesis should be at low levels among beginners, but should rise as skill is acquired. A second objective was to compare the extent of dependable all-senses and of nonvisual (i.e., kinesthetic) feedback for persons at various levels of typing skill. The third objective was to estimate the effects on typing performance

[1] This research was begun at Southern Illinois University, continued at The City University of New York, and completed there under a small grant from the United States Office of Education (OE-6-10-224) covering data analysis. The writer gratefully acknowledges the assistance of Donald M. Medley and John McClain, who served as statistical consultants.

(speed and errors) of deprivation of vision.

The rationale underlying the experimental procedures is that if a correct motion "feels" right, then an incorrect motion should "feel" wrong. Accordingly, as will be described, the basic datum is the proportion of errors of the kind amenable to kinesthetic feedback that were in fact sensed as wrong when working under visual and under nonvisual conditions. Whether "error detection" captures all that could or should be meant by kinesthetic feedback is unknown, but it is the defining process of the present investigation.

METHOD

Subjects

Of a total *N* of 266 typists, ranging in skill from 9 through 108 words per minute (wpm), all but 10 of the 224 subjects (*S*s) at skill levels between 9 and 74 wpm were students in 11 intact typing classes, taught by 8 different instructors, in 4 different high schools and colleges. Of the remaining 42 *S*s (at 75–108-wpm skill levels), most were employed typists, a few were typing teachers, and a few were finalists in a national contest for high school typing champions. The relevant population is one of levels of typing skill (as measured by gross stroking speed in ordinary copy work under normal conditions); and the sample data are in terms of skill level and not in terms of stage of training or amount of work experience.

Design

All *S*s typed under each of three conditions, in turn:

Condition 1: Normal, that is, under the conventional and familiar instructions to follow the copy word by word. This condition provided the basis for classification of *S*s according to skill level.

FIG. 1. Paperboard shield for typing under nonvisual conditions.

Condition 2: Under instructions to space once and retype instantly any word thought to have been mistyped, before continuing with the next word.

Condition 3: Under the same special instructions as for Condition 2, but deprived of visual reference to the typewriter or to the typescript produced.

These three work conditions are referred to hereafter as Condition 1 (Normal), Condition 2 (Visual, i.e., all-senses), and Condition 3 (Nonvisual, i.e., kinesthetic). The design is not orthogonal. The absence of a fourth condition (visual deprivation, but without instructions to retype in the case of sensed error) was due to the inability of teachers to make sufficient class time available to the investigator. Accordingly, this study does not provide a pure measure of the effects on performance of visual deprivation. Instead, Condition 2 furnished a measure of all-senses feedback and a means of accounting for that part of Condition 3 behavior attributable to the novel instructions to retype in the case of error. That is, Conditions 2 and 3 provided estimates of the effects of visual deprivation, holding constant the novel instructions to retype.

Test Content and Administration

Test copy consisted of three different sets of ordinary prose (approximately 875 words in each set), composed so as to be of equivalent difficulty as measured by syllabic intensity (mean number of syllables per word). Specifically, since a syllabic intensity of 1.40 is conventionally taken to represent average difficulty of copy materials for vocational typists, each successive 100 dictionary words of copy contained 140 syllables. Test copy was used in counterbalanced order across conditions within typing skill levels. Instructions to *S*s under all three conditions were to type at their ordinary rates, aiming at their best overall performance, giving due account to speed and accuracy. Under Conditions 2 and 3, *S*s were in addition urged to "try to 'catch' (as demonstrated by retyping) as many of the errors you do happen to make as you possibly can." Further, a sample of typescript containing every class of typing error, with appropriate retyping, was examined by *S*s before the work session, as an illustrative model. Also, to reduce novelty effects, *S*s were given a minute or two of unscored practice just before formal work under Conditions 2 and 3. Finally, with 10 exceptions, each *S* worked at his own accustomed typewriter.

For those *S*s who were students in training, Conditions 1 and 2 were administered in that order during one class period, with appropriate rest between the two. Condition 3 was administered in class the next day. For those *S*s who were employed typists, all three work conditions were administered the same day, in 1-2-3 order, and with appropriate rest between work sessions. Test-condition order was deliberately not counterbalanced in order (*a*) to maximize the reliability of the basis for classification of *S*s according to skill level and (*b*) to provide maximum experience with

retyping of errors prior to working under nonvisual conditions. Each of the three work conditions involved 12 continuous minutes of typing. To preclude interruption during the work session for changing paper in the machine, paper of sufficient length was cut from teletype rolls.

Apparatus

Deprivation of vision (under Condition 3) was accomplished by means of a paperboard shield, weighing 9½ ounces. As displayed in Figure 1, an adjustable neck halter and waist strap provided for individual differences in body build and in visual acuity, while holding the shield stably in place, yet entirely blocking S's view of the typewriter and of the typescript, but permitting free movement of the hands beneath the shield for operating the typewriter. Stimulus materials (printed in double column on an 8½ × 11-inch page) were tacked to the upper surface of the shield in any position desired by the typist. Routine inquiry of all Ss elicited their report that the shield was comfortable and stable and freely permitted operation of the typewriter.

Data and Analysis

Data consist of measures of speed (gross wpm) and of total errors under each of the three work conditions, as well as—under Conditions 2 and 3—of "applicable errors" (those amenable to kinesthetic feedback, excluding, e.g., incorrect word divisions or the omission of words or lines in the copy). Level of dependable feedback under visual (all-senses) and nonvisual (kinesthetic) conditions is defined as the percentage of "applicable errors" followed by immediate retyping, including the very occasional retyping of words that had not originally been mistyped. The 266 Ss were classified (on the basis of their speed under the normal work procedures of Condition 1) into ten 10-wpm skill ranges, but with unequal Ns in each skill range. To provide unbiased estimates of means for the entire range of typing skill, an equal frequency subsample of 189 Ss was drawn (21 from each of 9 skill ranges, combining 85–108-wpm typists into one cell).

RESULTS AND DISCUSSION

Results are presented in turn for (*a*) kinesthetic feedback, (*b*) kinesthetic versus all-senses feedback, and (*c*) effects on speed and errors of visual deprivation.

Dependable Feedback

With percentage of dependable feedback defined as the proportion of all errors amenable to kinesthetic feedback that were in fact followed by retyping under Visual and Nonvisual conditions, *M*s, *SD*s (in paren-

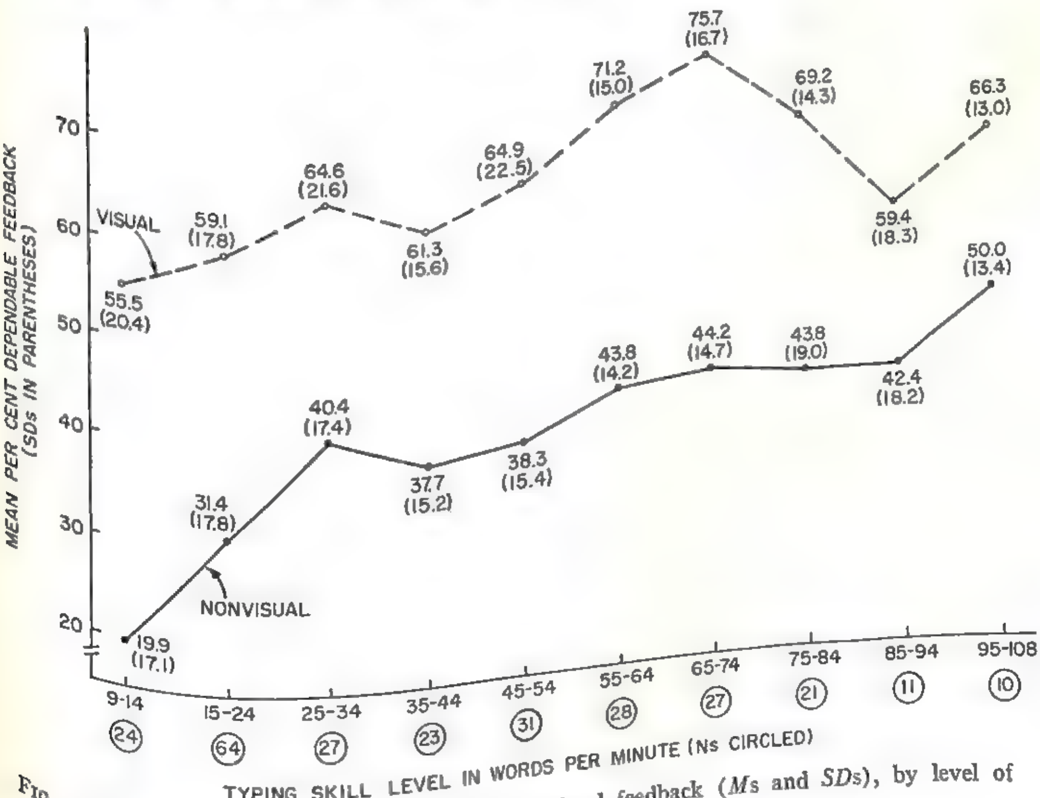


FIG. 2. Percentage of dependable visual and nonvisual feedback (*M*s and *SD*s), by level of typing skill, for 10-wpm skill ranges.

theses), and Ns (circled) for $N = 266$, distributed unequally among ten 10-wpm skill ranges, are displayed in Figure 2.

Descriptively, it is apparent from the data of Figure 2 that there was a sharp rise from low levels of kinesthetic feedback among novices. However, the long plateau thereafter was not anticipated and may or may not be characteristic of other skills. The surprisingly low (50%) levels of dependable kinesthetic feedback among 100-wpm typists contradicts and shows to be a delusion the common self-report by such typists that they "nearly always" know whether they make an error because it "feels" wrong. The skillful typist is apparently simply not aware of the frequency with which he steals corner-of-the-eye glances at keyboard and/or typescript as guidance for responses and for ascertaining their correctness or incorrectness.

One-way analysis of variance for percentage of dependable kinesthetic feedback (between and within 9 skill ranges for $N = 189$) resulted in an F of 5.283, for which $p < .001$ ($df = 8/180$). Tukey's significant gap test showed the differences in dependable feedback to lie between the 9–14-wpm typists and each of the other more advanced skill levels and between 15–24-wpm typists and those above 55-wpm levels ($p < .05$). No other significant interskill level differences were found. The hypothesis of a significant rise in dependability of kinesthetic feedback from low levels among beginners is thus confirmed; but it is equally clear that there is no further significant rise in utilization of kinesthetic cues once modest levels of typing skill have been reached. The curve for utilization of kinesthetic cues as a function of typing skill level is essentially negatively accelerated.

TABLE 1

ANALYSES OF VARIANCE FOR I. VISUAL AND NONVISUAL FEEDBACK, II. APPLICABLE ERRORS, III. TOTAL ERRORS, AND IV. SPEED

Source	df	I. Feedback		II. Applicable errors	
		MS	F	MS	F
Conditions (C)	1	3262.48	187.22*	937.63	62.14*
Range (R)	9	90.68	5.22*	19.28	1.28
C × R	9	17.13		14.45	
Error	512	17.37		15.09	
Total	531				
		III. Total errors		IV. Speed	
Conditions (C)	2	682.27	31.08*	105.75	131.69*
Range (R)	9	28.96	1.32	2515.66	3138.83*
C × R	18	23.13	1.05	2.18	2.72*
Error	768	21.95		.80	
Total	797				

Note.—$N = 266$, $R = 10$.
* $p < .001$.

TABLE 2

MEAN NUMBER OF APPLICABLE ERRORS UNDER VISUAL AND NONVISUAL WORK CONDITIONS (BY SKILL LEVEL)

Skill level (gross wpm)	N	M no. errors		Increase nonvisual-visual	
		Condition 2: Visual with retyping	Condition 3: Nonvisual with retyping	No. errors	% of errors
9–14	24	18.17	41.38	23.22	127.9
15–24	61	18.89	34.27	15.38	81.4
25–34	27	26.26	39.26	13.00	49.5
35–44	23	26.22	41.17	14.95	57.0
45–54	31	23.55	43.19	19.64	83.4
55–64	28	23.86	32.14	8.28	34.7
65–74	27	22.59	37.67	15.08	66.8
75–84	21	21.91	35.91	14.00	63.9
85–94	11	21.27	28.27	7.00	32.3
95–108	10	23.20	29.60	6.40	27.6
Total	266				
Grand M[a]	189	23.27	37.28	14.02	60.25

[a] These are based on a randomly selected 21 persons from each of 9 skill levels, combining 85–108 wpm into one cell.

Kinesthetic versus All-Senses Feedback

The losses in dependable feedback when typists were deprived of vision are directly observable in the differences between visual and nonvisual feedback percentages shown in Figure 2, namely, about a one-third loss for 9–14-wpm typists, about a one-fourth loss for 15–84-wpm typists, and about a one-sixth loss for 85–108-wpm typists. Analysis of variance for the data of Figure 2 is shown in Section I of Table 1.

As shown in Table 1 (Section I), condition effects and effects for each of the 10 skill ranges were highly significant. On the plausible hypothesis that differences between visual and nonvisual feedback levels ought to decrease with increases in skill, the present analysis shows no significant interaction between work condition and skill range. However, the absence of significant interaction is apparently due to the quite fine stratification of skill levels in 10-wpm ranges; for a parallel analy-

sis of variance for an equal frequency sample of 42 Ss drawn at random from each of four 25-wpm skill ranges (9–24, 25–49, 50–74, 75–108 wpm)—one that used Condition × Ss within Range as the error term—showed significant Condition × Range interaction ($p < .05$). The tendency toward narrowing visual-nonvisual feedback differences with increases in skill is somewhat uncertain: it can be found with broad, but not with narrow, grouping according to skill range.

Effects of Work Conditions on Speed and Errors

Results are presented in turn for applicable errors (those amenable to kinesthetic feedback), total errors, and speed.

Applicable errors. Applicable error means for each skill range and differences in means as between visual and nonvisual conditions are shown in Table 2.

Applicable errors, as is apparent from comparison with total error data (Table 3), made up more than 90% of all errors; most typing errors are amenable to kinesthetic feedback. As shown in Table 2, visual deprivation was accompanied by very substantial (60%) increases in such errors. Specifically, the extreme right-hand column of Table 2 suggests classification of typists into three groups with respect to effects of visual deprivation on errors susceptible to kinesthetic feedback: very large negative effects on typists at 9–24-wpm levels, intermediate effects on 25–84-wpm typists, and markedly smaller (but decidedly nontrivial) effects on 85–108-wpm typists.

Analysis of variance for applicable errors is shown in Section II of Table 1. As with the feedback analysis of Section I, condition differences for applicable errors were highly significant. The absence of significant range effects (and of interaction) is implicit in the essentially "plateau-ish" character of the differences in condition means for those across the broad 25–84-wpm range mentioned earlier (extreme right-hand column of Table 2).

Speed and total errors. Means for speed and total errors under each of the three work conditions and differences between means for visual and nonvisual work conditions are shown in Table 3 for the equal-frequency sample of 189 Ss.

TABLE 3

SPEED AND TOTAL-ERROR MEANS UNDER EACH OF 3 WORK CONDITIONS AND ERROR INCREASE UNDER NONVISUAL CONDITIONS

Variable	Condition 1: Normal	Condition 2: Visual with retyping	Condition 3: Nonvisual with retyping	Error increase nonvisual-visual No.	Error increase nonvisual-visual %
Speed (wpm)	50.30	45.30	44.46		
Total Errors	39.27	25.74	40.10	14.36	55.8

Note.—$N = 189$.

Although total error means for each of the 10 ranges individually are not shown in Table 3, they paralleled in trend those for applicable errors shown in Table 2, as did the percentage of increase in total errors under nonvisual conditions (e.g., 114% increase for 9–14-wpm typists down through a 30% increase for 85–108-wpm typists). Concerning speed, the 1.04-wpm decrement covers differences for each of the 10 individual skill levels ranging from −.48 to 3.43 wpm.

Analysis of variance for total errors is shown in Section III of Table 1. Following the finding of significant overall conditions effects, the differences were found to lie between Conditions 1 and 2 and between Conditions 2 and 3 ($p < .001$). The reduction in errors brought about by instructions to retype was nearly equal to the increase in errors accompanying subsequent deprivation of vision; accordingly, Conditions 1 and 3 did not differ significantly.

Analysis of variance for speed is shown in Section IV of Table 1. It is apparent from the speed means shown in Table 3 that instructions to retype had a larger (negative) effect on speed than did deprivation of vision. Thus, following the finding of significant overall conditions effects (Table 1, Section IV), significant differences were found between Conditions 1 and 2 and between Conditions 1 and 3 ($p < .001$), but not between Conditions 2 and 3.

Correlational Evidence

Pertinent correlational data are in substantial agreement with and therefore support

the early data on means. For the equal-frequency subsample of 189 *Ss*, $r = .36$ was found between typing skill (gross wpm under normal work conditions) and percentage of dependable nonvisual feedback (Condition 3). There was not any strong tendency for continuous increase in dependable kinesthetic feedback as typing skill increased. The sharp effects of visual deprivation at very low skill levels and the modest increase in kinesthetic dependability among the very fastest typists were swamped by the general absence of differences in effects of visual deprivation among typists in the great middle range of 25–94-wpm speeds. Another factor (as shown by the *SD*s in Figure 2) was the overlap in levels of dependable feedback between skill levels. Finally, $r = .39$ was found for percentage of dependable feedback between Conditions 2 and 3. The modest size of this r (in terms of r^2) suggests the substantial independence of kinesthetic feedback from the other components of all-senses feedback (mainly visual, but occasionally tactile and auditory).

Implications for Training

The results of this study call into question the conventional insistence on "touch" typewriting from the start of learning and suggest the desirability of free use of vision early in learning to typewrite: as guidance for making responses, as a source of corrective information for wrong responses and of immediate reinforcement for correct responses, and as a reducer of the anxieties and tensions that reportedly commonly accompany early insistence on nonvisual work.

REFERENCES

CARR, H. A. The influence of visual guidance on maze learning. *Journal of Experimental Psychology*, 1927, **4**, 399–417.

FITTS, P. M. Engineering psychology and equipment design. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 1287–1340.

FLEISHMAN, E. A., & RICH, S. Role of kinesthetic and spatial-visual abilities in perceptual-motor learning. *Journal of Experimental Psychology*, 1963, **66**, 6–11.

HONZIK, C. H. The sensory basis of maze learning in rats. *Comparative Psychology Monographs*, 1936, **13**, No. 64.

KOCH, H. L., & UFKESS, J. A comparative study of stylus maze learning by blind and seeing subjects. *Journal of Experimental Psychology*, 1926, **9**, 118–131.

ZIGLER, M. J. Touch and kinesthesis. *Psychological Bulletin*, 1932, **29**, 260–278.

(Received April 28, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 167-169

INDIVIDUAL RESPONSES AND SOCIAL DESIRABILITY

CHARLES HANLEY

Michigan State University

Several investigations indicate that, for the average S, endorsement of personality inventory items has only a low to moderate correlation with social desirability (SD), unlike the case when probability of endorsement by a group is related to desirability. These studies have used point-biserial correlation for the individual and the product-moment coefficient for the group. When a single S's responses are analyzed in a way that permits estimation of probability of endorsement, the product-moment correlation between this probability and SD typically is high, approaching much closer to the group value. High average individual correlation between endorsement and desirability does not prevent a diagnostic scale from discriminating between disparate groups.

As a general rule, the probability that members of a group will endorse a personality inventory item correlates highly with independent ratings of the social desirability (SD) of the item. Yet the picture appears to change with responses and ratings of individual subjects (*S*s). Despite differences in samples and procedures, Taylor (1959), Boe and Kogan (1963), Block (1965), and Edwards (1965), all report endorsement-SD correlations much lower in individuals than in groups. All have used the point-biserial coefficient for the individual and the product-moment for the group. The discrepancy between individual and group results leads Block to observe: "The use of a solitary coefficient based on grouped data clearly is erroneous in its suggestion as to the potency of SD in influencing item endorsements in the individual case [p. 75]." Boe and Kogan, as well as Taylor, reach a similar conclusion. Still, these investigators have had to consider two methodological factors as contributing to the discrepancy: (*a*) unreliability of individual endorsements, and (*b*) limiting properties of the point-biserial coefficient.

Boe and Kogan (1963) found by retest a mean individual endorsement reliability of .77. Correcting the mean individual endorsement-SD correlation for attenuation raised it only from .27 to .32, compared to a .53 group coefficient. (All endorsement-SD correlations in their study are lower than usual for the MMPI, because the SD values of their sample of items deviate widely from the total distribution of MMPI SD values.)

Taylor (1959) briefly, and Block (1965) more fully, treated the reliability question hypothetically. Taylor estimated mean individual endorsement reliability as .60, a value that now appears low, and calculated the mean individual endorsement-SD correlation of .39 in his study would rise to .60 if unreliability were removed. Block considered individual endorsements highly reliable, citing Goldberg and Rorer's (1963) finding of about 85% mean individual retest consistency on the MMPI. Simple computations bear out Block's conclusion that the difference between his mean individual coefficients, which approximate .54 in five different samples, and corresponding group coefficients, which average near .81, cannot be attributed to attenuation, for when "true" and "false" responses of an individual *S* split 40/60 (or vice versa), 85% consistency of response yields a tetrachoric correlation of around .86 between test and retest. For true/false splits between 30/70 and 50/50, the correlation must be similarly high to produce 85% response consistency. Since Block used SD values which Messick and Jackson (1961) reported to correlate .96 with a separate set developed earlier, unreliability either in endorsement or in SD cannot account for the gap between individual and group coefficients.

The second methodological problem, limiting characteristics of the point-biserial, is treated by Block (1965) and Edwards (1965). Block pointed out that

> Point-biserial *r*s, although product-moment coefficients, fall on a more compressed scale than

product-moment *rs* based solely on continuous variables. . . . Moreover, biserial coefficients are further lowered as an individual's proportions of "trues" and "falses" deviates from 50–50 [p. 76].

Edwards developed further the consequences of these aspects of the point-biserial; if an *S* gave the socially desirable response to each of the items used by Boe and Kogan (1963), the point-biserial correlation between endorsement and SD would be .81, not 1.00. Making an additional assumption about the negligibility of small differences between ratings that fall within the neutral sector of the SD scale, Edwards calculated that the maximum possible individual coefficient for Boe and Kogan's items is approximately .62. It seems quite clear that comparisons of point-biserial and product-moment coefficients in the present application must typically show wide discrepancies.

Fortunately, it is feasible to assess the endorsement-SD relation in the individual *S* by a product-moment coefficient. Giving up some precision in measuring SD—precision that may be superfluous in most practical instances—makes it possible to use a single *S*'s responses to subsets of items, each subset relatively homogeneous in SD, to compute a more reliable estimate of probability of endorsement than is provided by single item data. Consider, for example, the case where SD values of items range between 1.00 and 9.00, as in the Messick and Jackson (1961) norms for the MMPI. All items with SD values between 1.00 and 2.00 can be pooled, and probability of endorsement calculated from an individual's responses to the pool. The same can be done with items falling between 2.00 and 3.00 on the SD scale, and so on for the remainder. The pools arrived at in this manner, however, need further subdivision into subscales of equal length if: (*a*) each point on the scattergram relating endorsement and SD is to be based on the same number of items, and (*b*) the scattergram is to have enough points to be roughly comparable to the plot obtained when group endorsement is related to SD for each item.

The following use of this alternative method shows that the tendency of individuals to give socially desirable responses to MMPI items is much closer to that prevailing for the group than earlier investigations have found with point-biserial coefficients. Because *Ss* in this demonstration come from dissimilar groups, it also yields information regarding the proposition that a high individual endorsement-SD relation will prevent discrimination by diagnostic scales.

METHOD

To simplify scoring, the demonstration employed the first 300 items on the standard MMPI booklet answer sheet. Items first were grouped, using the Messick and Jackson norms, into eight desirability classes: 1.00–1.99, 2.00–2.99, . . . , 8.00–8.99. This grouping gives a skew distribution, the mode in the category 3.00–3.99. Practical considerations set the length of the subscales; subscales of six items each used most of the items and minimized losses at the extremes of the SD dimension. Subscales preserved the order in which items appear in the MMPI booklet; for example, the first six booklet items with SD values between 1.00 and 1.99 made up the first subscale at the level, the second six the second subscale, and so on until no more scales could be formed at that SD level. Use of this procedure yielded 46 subscales, each relatively homogeneous in SD, for a total of 276 items. These items constitute the majority of those on the standard MMPI scales, and probably involve more extreme SD values than the inventory as a whole.

The 46 subscales were scored for each *S*. Because every subscale had six items, probability of endorsement could be represented by number of items endorsed per scale. For each *S*, scores (ranging from 0 to 6) on the 46 subscales were correlated by the product-moment coefficient with their SD levels (ranging from 1 to 8). The same data were analyzed by point-biserial correlation. Finally, responses pooled across *Ss* gave group probability of endorsement, which then was correlated with SD. In all instances, SD was represented by the grouped values, a procedure that kept errors, due to roughness of grouping, constant over all three methods.

Subjects. MMPI protocols came from two disparate groups. First were those of 30 male undergraduates in introductory psychology at Michigan State University in 1954, collected in an unrelated study. Second were 30 protocols of convicts from routine testing upon entry into Southern Michigan Prison in 1959.

Results. Analysis of responses by point-biserial coefficients gives results like those in earlier studies. For students, point biserials ranged from .18 to .67 with a mean (computed by Fisher's *z* transformation) of .51. The group correlation was .93. For prisoners, point biserials ranged from .16 to .68 with a mean of .46; the group product-moment coefficient again was .93. As expected, there are wide discrepancies between apparent individual and group tendencies to give socially desirable responses.

These discrepancies are much narrower with individual product-moment correlations. For students, individual coefficients ranged from .38 to .91 with a mean of .80. Coefficients for prisoners ranged from .30 to .90 with a mean of .77. As might be anticipated, the product-moment coefficient always was higher than the corresponding point-biserial. (Biserial coefficients fell between the other two types.)

Past studies lead to the expectation that the prisoners would have higher *Pd* scores than the college men. Among the 30 prisoners, 27 had *T* scores greater than 70, as compared to 3 of the college men, this being true both for *K*-corrected and uncorrected scores. Contrasted to this high concurrent validity for the *Pd* scale, the individual endorsement-SD coefficients almost completely overlap in the two groups.

CONCLUSIONS

Point-biserial coefficients for the present *S*s indicate that about 25% of individual endorsement variance is associated with SD, but the product-moment method of analysis shows that 60% is a more likely figure. Clearly the tendency of individuals in these two different samples to give answers judged elsewhere as socially desirable is strong rather than weak. Because these responses, when analyzed by the point-biserial coefficient, have about the same apparent relation to SD as earlier studies reported from use of a similar technique, it is hard not to conclude that they underestimated the magnitude of the relation that actually exists.

The high SD-endorsement correlation in the typical individual protocol does not prevent the most relevant diagnostic scale, *Pd*, from discriminating between prison and college samples. Earlier discussions of SD and personality inventories sometimes have seemed to argue that if SD correlates highly with endorsement, there is no point to the use of separate diagnostic scales; or, contrariwise, that if separate diagnostic scales are valid, there cannot be a high correlation between endorsement and SD in individuals. Results from the present study show plainly that a strong association between individual endorsement and SD can exist alongside of high concurrent validity for a diagnostic scale, a finding consistent with a recent study by Wahler (1965) demonstrating that reliable item preferences distinguish groups when the effects of SD are partialed out.

Neither exaggerating nor underestimating the presence of SD in personality inventories serves a useful purpose. The relation of endorsement to SD has basic interest apart from the fact that it occurs in inventories originally designed for applied purposes. If practical considerations are the center of interest, then SD surely is involved in the design of suppressor scales and of disguised or "subtle" items.

REFERENCES

BLOCK, J. *The challenge of response sets.* New York: Appleton-Century-Crofts, 1965.

BOE, E. E., & KOGAN, W. S. Social desirability response set in the individual. *Journal of Consulting Psychology,* 1963, **27**, 369.

EDWARDS, A. L. Measurement of individual differences in ratings of social desirability and in the tendency to give socially desirable responses. *Journal of Experimental Research in Personality,* 1965, **1**, 91–98.

GOLDBERG, L. R., & RORER, L. G. Test-retest item statistics for original and reversed MMPI items. *Oregon Research Institute Research Monograph,* 1963, **3**, No. 1.

MESSICK, S., & JACKSON, D. N. Desirability scale values and dispersions for MMPI items. *Psychological Reports,* 1961, **8**, 409–414.

TAYLOR, J. B. Social desirability and MMPI performance: The individual case. *Journal of Consulting Psychology,* 1959, **23**, 514–517.

WAHLER, H. J. Item popularity and social desirability in the MMPI. *Journal of Applied Psychology,* 1965, **49**, 439–445.

(Received April 30, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 170–173

ORDER EFFECTS IN ASSESSMENT DECISIONS [1]

PATRICIA M. ROWE

University of Waterloo, Ontario

This study was concerned with the effects of order of presentation on judgments of people. 3 groups of 25 Ss each judged descriptions of individuals containing varying proportions of favorable and unfavorable adjectives. 3 orders of presentation were used: ascending (all unfavorable adjectives through to all favorable), descending, and random. More descriptions were accepted with the ascending order than with the descending order. Furthermore, order of presentation significantly affected individual differences in category width. The results were interpreted in terms of assimilation and contrast notions.

Most of the personnel-selection literature treats an assessment decision as if it were a single, isolated judgment. Yet most selection decisions, whether they concern job applicants or candidates for graduate work, are made in the context of previous, similar judgments. There is a wide variety of possible contextual effects on decisions: the type and number of previous judgments, the time interval between judgments, and the order in which judgments are made, to mention only a few. In a study examining the effect of one judgment on the following judgment, Holmes and Berkowitz (1961) found that an individual was judged more favorably when he was preceded by an extremely unfriendly stimulus person than when he was preceded by a very friendly stimulus person. The present investigation is concerned with the effects of making a series of judgments on an ordered array of stimulus persons. Simply put, the question here is whether decisions are differentially influenced if one happens to judge applicants in order from best to worst, or from worst to best.

Many previous studies of judgments about people have been concerned not with the ordering of individuals, but rather with the ordering of characteristics within a single individual. Under these latter conditions, studies by Anderson and Barrios (1961), Bolster and Springbett (1961), and Luchins (1957) have demonstrated primacy effects: the characteristics presented first have a greater effect on the total impression of an individual than do characteristics presented later. Berkowitz (1960) has argued that such primacy effects are due to the assimilation of later material into the earlier impression. Assimilation occurs because all the characteristics belong to the same individual (i.e., the psychological distance between early and late materials is small). If, however, subjects (*S*s) recognize the incompatibility of the two sets of material, the psychological distance is increased, and a recency effect due to contrast is likely. Under the condition of a systematically ordered group of individuals, where the sets of material appear in different individuals, judgments would be expected to show the effects of contrast. Thus, a person should be judged more favorably when preceded by a less likable person than when preceded by a more likable person. Such a prediction is identical to the one that would be made from Helson's (1964) adaptation-level theory, and has already been shown applicable to psychophysical judgments (Helson & Nash, 1960).

Because judgments of other people are here regarded as a categorizing process, it is also necessary to consider individual differences in categorization. The writer has shown that one type of individual differences, category width (the number of stimuli placed in the same category or class), exists in judgments of people. When personnel officers make judgments of written descriptions of job applicants they show large and consistent differences in the number of applicants selected (Rowe, 1963). Furthermore, Bruner and

[1] This study was supported by Defence Research Board of Canada Grant No. 9425-10 to D. Bindra of McGill University and Grant No. 9401-27 to the writer. Thanks are due to M. P. Bryden for his many helpful comments.

Tajfel (1961) have demonstrated that individual differences in category width in a dot-classifying situation are related to sensitivity to change in sequentially presented stimuli. These results suggest that individual differences in category width may be affected by systematically ordered stimuli, a question which will be examined in the present study.

METHOD

Subjects

Students in the extension (noncredit) section of an introductory psychology course at McGill University and army recruits served as *Ss*. The distribution of sex, age, and educational level in the sample was comparable to that found in the general adult population; since no differences as a function of any of these variables were found, no further reference to these characteristics of the *Ss* will be made. A total of 82 *Ss* were tested.

Stimuli

A set of 299 adjectives was presented to 50 pilot *Ss* for judgments of "favorable," "unfavorable," or "neither favorable nor unfavorable." Out of this set 30 favorable (F) and 30 unfavorable (U) adjectives were selected, with the restrictions that at least 90% of the *Ss* had agreed on their ratings, and that any two adjectives could coexist in the same description (i.e., no adjective was the antonym of any other adjective in the set). Descriptions of individuals were then constructed by selecting six of these adjectives. A total of 103 descriptions were constituted in the following fashion: seven sets of nine descriptions each, one set of each of the Type 6F adjectives, 5F-1U, 4F-2U, 3F-3U, 2F-4U, 1F-5U, and 6U; and 40 additional descriptions containing 3F and 3U adjectives. Within each description the adjectives were arranged in random order. A sample description of the 3F-3U type is:

Individual X is quarrelsome, dissatisfied, logical, organized, tactless, and capable.

Procedure

The 103 descriptions were arranged in three different orders. For all *Ss* the first 30 descriptions, of the 3F-3U type, were the same. Similarly, the last 10 descriptions, again of the 3F-3U type, were the same for all *Ss*. Descriptions 31–93 constituted the experimental treatment. In the ascending (A) order Descriptions 31–39 were of the 6U type, followed by the nine Type 1F-5U descriptions, and so on through to the Type 6F descriptions. In the descending (D) order the experimental treatment started with Type 6F descriptions and finished with 6U descriptions. In the random (R) order each consecutive block of seven descriptions contained one description of each type, randomly arranged within blocks. The 103 descriptions were presented to *Ss* in typed test booklets without any divisions between the initial set (Descriptions 1–30), the experimental set (Descriptions 31–93), and the final set (Descriptions 94–103). Previous work had found that, from the point of view of decision processes, there is great similarity in judgments, whether *Ss* are asked to make decisions about job applicants, prospective roommates, or friends. Because it was more realistic for *Ss* in this study, their task was to read each description, decide whether they would like to have this person as a friend, and place a check mark in the appropriate space beneath the description. The *Ss* were instructed to read each description in turn, starting with the first description, and make their decision before going on to the next description.

The score for the various parts of the questionnaire was the number of "like" judgments. Thus each *S* was scored for the frequency of his like responses on the initial set, the final set, and on each of the seven types of experimental descriptions.

Since a previous study (Rowe, 1963) had found large individual differences in judgments of this sort, it was necessary to equate the three groups in terms of their number of like judgments. The score on the first 30 descriptions was used for this purpose. After discarding seven *Ss*, the mean number of like responses for the three orders on Descriptions 1–30 were: A, 11.2; D, 10.7; and R, 10.8. The standard deviations were: A, 4.2; D, 3.7; and R, 4.2. The number of *Ss* per order was 25.

RESULTS AND DISCUSSION

The effects of order of presentation on the number of like decisions can be seen in Figure 1. Observe that with Order A more descriptions are liked than with Order R through most of the experimental treatment, and that with Order D fewer descriptions are liked than with Order R. These data were submitted to an analysis of variance

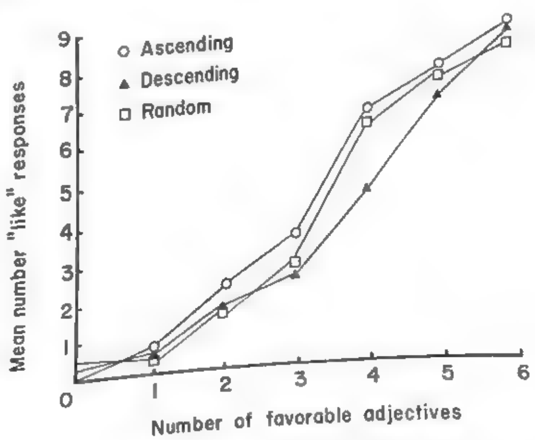


FIG. 1. Mean number of "like" responses as a function of number of favorable adjectives for the three orders of presentation.

TABLE 1

ANALYSIS OF VARIANCE OF LIKE RESPONSES ON DESCRIPTIONS 31–93

Source	df	MS	F
Between subjects			
Order (O)	2	22.09	3.93*
Error between subjects	72	5.62	
Within subjects			
Description type (D)	6	788.88	419.62**
O × D	12	5.49	2.92**
Error within subjects	432	1.88	

* $p < .05$.
** $p < .01$.

(Order × Type of Description, Lindquist, 1953, mixed design Type I), the results of which appear in Table 1. The main effect of order of presentation is significant: the means of the A and D orders for Descriptions 31–93 are significantly different from each other. Thus the earlier prediction has been confirmed—stimulus persons systematically ordered in terms of a favorable-unfavorable dimension produce a contrast effect on judgments of like or dislike of these persons. In other words, if one first encounters an unpleasant individual in a social situation he is more likely to regard an average person favorably than if he first meets a very pleasant individual. These results are also further confirmation for the notion that the underlying variable responsible for producing assimilation and contrast is that of psychological distance—the greater this distance, providing the stimuli are still on the same continuum, the more likely is contrast. In the present study, psychological distance was great because the stimuli were separated into different individuals, and thus contrast resulted.

Type of description, of course, produced a highly significant effect on the number of like decisions. It is perhaps of interest to note in Figure 1 that with all conditions, even with Order R which is uncontaminated by order effects, the number of like judgments does not increase linearly with the number of favorable adjectives. Most particularly, it can be observed that the addition of one favorable adjective from Type 3F-3U to Type 4F-2U produces a greater increase in the number of like responses than does the addition of one favorable adjective from Type 2F-4U to Type 3F-3U. Though beyond the scope of the present paper, it appears that there is some maximum number of unfavorable traits permissible in a friend; above that number, most individuals are rejected, while below it, most individuals are accepted.

The significant interaction between order of presentation and type of description can best be understood by reference to Figure 1. The largest differences among the three orders occur with the "medium" descriptions, those containing one or more of both favorable or unfavorable adjectives. Though such a finding might be due to ceiling and floor effects, its similarity to results obtained by Ball (1953) and Hovland and Sherif (1952) suggests a different interpretation. The latter investigators found that statements around the center of attitude scales are more readily shifted than are statements farther from the center. Thus the significant interaction in the present study is probably due to the greater susceptibility of average persons to order effects.

The purpose in presenting Descriptions 94–103, all of Type 3F-3U, was to determine whether there are any long-term effects of order of presentation. In Table 2 the mean percentage of possible like responses for Descriptions 1–30 and 94–103 is given for each of the three orders. From this table it is clear that there are no differences among the three orders on Descriptions 94–103, nor

TABLE 2

MEANS AND CORRELATIONS OF PERCENTAGE OF LIKE RESPONSES ON DESCRIPTIONS 1–30 AND 94–103 AS A FUNCTION OF ORDER OF PRESENTATION

Order	M percentages		Correlations between Descriptions 1–30 and 94–103
	Descriptions 1–30	Descriptions 94–103	
A	37.47	37.60	.546*
D	35.60	36.40	.269
R	35.87	35.60	.712*

* $p < .01$.

is there any change in the percentage of like decisions from Descriptions 1–30 to 94–103. Although these results suggest that there are no long-term group effects from order of presentation, they do not answer the question of whether systematic stimulus orders affect individual differences in decision behavior. In the last column of Table 2 the correlations between number of like decisions in Descriptions 1–30 and 94–103 are presented. Individual differences in the number of like judgments are very consistent from the initial to the final set under the random order, but are considerably disrupted under both the A and D orders. Moreover, the correlation for Order R is significantly greater than that for Order D ($p < .05$). It seems that category width is a reliable phenomenon only under the condition of randomly presented stimuli. Thus the process underlying category width may well be associated with responsivity to a random assortment of stimulus items. Exposing *S* to a systematic order of stimuli tends to change his responsivity in a direction not predicted by initial category width. Such a notion would not be incompatible with Bruner and Tajfel's (1961) suggestion that category width is related to sensitivity to stimulus change.

Studies of this sort are difficult if not impossible to carry out with live people instead of descriptions as stimuli (for a discussion of this point see Webster, 1964, p. 72). Thus only through studies such as the present one can one evaluate the effects of order in the selection situation. The effects of previous judgments on later ones are quite striking and present clear evidence for the contribution of the context to assessment-decision variance. Whether an individual is accepted or rejected for a job may well depend more on the characteristics of the previous applicants than on his own traits. Perhaps more important is the evidence concerning the effects of ordered presentations on individual differences in decision behavior. If one considers category width, or the number of like responses, a measure of an interviewer's standard of acceptance of job applicants, it would seem to be a serious mistake to make judgments about applicants in other than a random order. Otherwise, if judgments about applicants are made sometimes in ascending order and sometimes in descending order, the interviewer's standard of acceptance will show gross fluctuations from one judgment session to another.

REFERENCES

ANDERSON, N. H., & BARRIOS, A. A. Primacy effects in personality impression formation. *Journal of Abnormal and Social Psychology*, 1961, 63, 346–350.

BALL, J. H. The influence of background and residual stimuli upon the measurement of attitudes. Unpublished master's thesis, University of Texas, 1953. Cited in H. Helson, *Adaptation-level theory*. New York: Harper & Row, 1964.

BERKOWITZ, L. The judgmental process in personality functioning. *Psychological Review*, 1960, 67, 130–142.

BOLSTER, B. I. I., & SPRINGBETT, B. M. The reaction of interviewers to favorable and unfavorable information. *Journal of Applied Psychology*, 1961, 45, 97–103.

BRUNER, J. S., & TAJFEL, H. Cognitive risk and environmental change. *Journal of Abnormal and Social Psychology*, 1961, 62, 231–241.

HELSON, H. *Adaptation-level theory*. New York: Harper & Row, 1964.

HELSON, H., & NASH, M. C. Anchor, contrast, and paradoxical distance effects. *Journal of Experimental Psychology*, 1960, 59, 113–121.

HOLMES, D. S., & BERKOWITZ, L. Some contrast effects in social perception. *Journal of Abnormal and Social Psychology*, 1961, 62, 150–152.

HOVLAND, C. I., & SHERIF, M. Judgmental phenomena and scales of attitude measurement: Item displacement in Thurstone scales. *Journal of Abnormal and Social Psychology*, 1952, 47, 822–832.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

LUCHINS, A. S. Primacy-recency in impresson formation. In C. I. Hovland (Ed.), *The order of presentation in persuasion*. Vol. 1. New Haven: Yale University Press, 1957. Pp. 33–61.

ROWE, P. M. Individual differences in selection decisions. *Journal of Applied Psychology*, 1963, 47, 304–307.

WEBSTER, E. C. *Decision making in the employment interview*. Montreal: Industrial Relations Centre, McGill University, 1964.

(Received April 30, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 174-180

AN EMPIRICAL APPROACH TO CRITERION SPECIFICATION IN GENERAL BUSINESS[1]

JAMES E. PALLETT[2] AND DONALD P. HOYT

Placement Center, University of Iowa — American College Testing Program

The investigation sought to discover specific behavioral characteristics related to judgments of success in general business activities. Supervisory ratings of 23 specific characteristics and of 2 overall measures of success were obtained for 230 University of Iowa graduates employed in nonspecialized aspects of business. All Ss had been out of college for 5–10 yr. Multiple-regression analyses showed that 13 of these ratings were sufficient to account for the entire set of variances and covariances; 8 of these 13 ratings made independent contributions to the prediction of overall ratings and were thus considered to be elements of success in general business. Implications were drawn for job recruitment, selection, guidance, and training.

Most college students want to know "Do I have what it takes to be successful in this field or that field? If not, can I 'improve' myself so that I will be successful? How?" These questions are not much different than those asked by the recruiter of college talent. "Can this individual help us? In what division can his talents best be employed? What type of training will benefit him?"

The research reported in this paper seeks to define the elements of success in general (nontechnical) occupations in business. Its purpose is to describe relatively specific characteristics associated with a positive evaluation by employers.

Because different jobs require different skills, the study is limited to one type of employment, "general business." Of concern is that group of college graduates for whom specific or technical academic backgrounds are irrelevant to their employment. This includes graduates of business or liberal arts curricula whose employment in the business world bears little relationship to their college major. Examples include mathematics majors employed as office supervisors, political science majors employed as advertisers, or general business majors employed as salesmen. Excluded were such groups as art majors working as commerical artists, accounting majors working as accountants, or mathematics majors working as actuaries.

Sample

The sample was chosen from the placement list of seniors registered at the University of Iowa Business and Industrial Placement Office between 1954 and 1959. All had been out of college for from 5 to 10 years.

Random samples of 250 liberal arts and 250 business administration male registrants were taken from these placement lists. These 500 individuals were then contacted by postcard in order to learn the nature of their present employment and the name and address of their immediate supervisor. They were also asked if they had pursued graduate work.

A total of 415 of the 500 *Ss* (83%) returned completed postcards. Of these, 206 were in the liberal arts sample and 209 in the business administration sample.

The *Ss* were excluded for the following reasons: employed in a specialized job which required their academic training; earned an advanced degree; employed in government, education, or other nonprofit setting; self-employed or employed in family business. These exclusions reduced the sample size to 256 (103 liberal arts graduates and 153 business administration graduates).

Immediate supervisors of these 256 individuals were asked to supply ratings on qualities described below. Returns were received for 92 of the liberal arts graduates (81%) and for 138 of the business administration graduates (90%). These 230 *Ss* constituted the basic sample.

Measurement

To establish the general measure of success which these college graduates had attained, supervisors

---

[1] This paper is based on the doctoral dissertation of the senior author (Pallett, 1965), done under the direction of the junior author. The project was supported in part by a grant-in-aid from the Research Committee of the Midwest College Placement Association.

[2] Now at the University Placement Service, University of Wisconsin, Milwaukee.

were asked to rate their "progress" (advancement to date) and "potential" (anticipated advancement) in relation to others of similar age, experience, and tenure. Ratings were made on a 7-point scale ranging from "lowest 2%" to "highest 2%." While the two ratings were expected to overlap considerably, both were retained on the chance that some individuals would advance rapidly and soon reach their limits while others would be "brought along" gradually but hold promise for eventual promotion to very high leadership posts.

To assess the more specific characteristics believed to represent aspects of success, a series of 23 rating scales were constructed. These 23 characteristics represented a synthesis of specific characteristics commonly used in industry,[3] or commonly mentioned by vocational psychologists (Super, 1957; Super & Crites, 1962) and placement specialists (Dickenson, 1955; Johnson, 1959). The characteristics were named Adaptability, Creativity, Ability to Deal with Abstracts, Problem-Solving Ability, Judgment, Foresight, Decisiveness, Accuracy, Identification with the Business World, Identification with the Company, Dependability, Self-improvement, Interpersonal Relations, Sensitivity to Others, Likableness, Poise, Administrative Ability, Written Communication, Oral Communication, Leadership, Enthusiasm (drive), Persuasiveness, and Competitiveness.

Each scale consisted of three statements descriptive of behavior relevant to a given characteristic. One statement was favorable (scored 2), one neutral (scored 1), and one unfavorable (scored 0). The three statements in each set were arranged in a random order. The focus on behavior was intended to sharpen the meaning of the variables and to improve the reliability of ratings (Thorndike & Hagen, 1961). Two examples are given below:[4]

Decisiveness

a. He is able to make decisions quickly and firmly when called upon to do so.
b. When it is necessary for him to make a decision, he is able to do so with only ordinary qualms.
c. He tends to vacillate a great deal when faced with a decision that he must make.

Enthusiasm (drive)

a. He exhibits little enthusiasm for his work.
b. He is so involved in his work that he frequently works overtime voluntarily.
c. He carries on his duties with a normal amount of energy and enthusiasm.

PROCEDURE

Two goals guided the statistical analysis: (1) To determine if fewer than 23 ratings could be used to represent the universe of behaviors descriptive of performance in general business activities; (2) To determine which of the 23 ratings made independent contributions to the prediction of overall success and thus could be considered "elements" of success in general business.

A factor-analytic approach was considered but rejected, for reasons suggested by Bechtoldt (1962). Briefly, the factor approach involved three difficulties. (*a*) In the present study, there was a clear status difference between the overall criterion ratings (dependent variables) and the specific ratings (which could be treated as either dependent or independent variables). Factor analysis is poorly suited to problems where such status differences exist. (*b*) It seemed unlikely that the data of this investigation could satisfy the rigorous conditions required before factor-analysis techniques can usefully be employed to provide objective definitions. (*c*) Factor-analytic procedures have been regarded traditionally as appropriate to exploratory or confirmatory purposes, but not to definitional purposes such as those guiding the present investigation.

The multiple-regression model obviates these difficulties and thus was chosen as a more appropriate approach to the problem.[5] The procedure[6] required selecting a few of the specific scales as independent variables, and treating all remaining scales as dependent variables. Then a matrix was prepared of the partial intercorrelations among the dependent variables with the independent variables held constant. The goal was to find a set of independent variables which accounted for the variances and covariances of all remaining dependent variables. An appropriate statistical test is given by Anderson (1958) for determining if the partial correlations, as a set, differ from zero; that is, the statistical hypothesis is that the partial correlations are independent.

Assuming that a set of independent variables could be found such that this hypothesis was tenable, the next step was to determine which of those variables made independent contributions to the prediction of the overall criteria. Such variables would constitute "elements" of success in general business.

RESULTS

Intercorrelations among the 23 specific ratings and the two overall ratings are shown in Table 1.

[3] Thanks are due to the following companies which allowed the authors to examine their personnel evaluation forms: The Maytag Company, General Motors Corporation, Caterpillar Tractor Company, American Hospital Supply Corporation, General Electric Company, Ford Motor Company, and Sears, Roebuck, and Company.

[4] Copies of the complete set of scales are available from the senior author.

[5] The authors are indebted to Harold Bechtoldt for providing suggestions regarding statistical analysis and for making available an appropriate computer program. Thanks are also due to Bill Snider and the University of Iowa Computer Center for providing data analysis services.

[6] A more detailed statement of the procedures is contained in Bechtoldt, Benton, and Fogel (1962).

TABLE 1

INTERCORRELATIONS OF 25 MEASURES OF JOB PERFORMANCE

Measures	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Adaptability	35	24	29	34	22	28	11	50	20	30	31	26	16	24	45	42	14	35	39	21	44	37	51	47
2. Creativity		43	39	36	47	38	25	41	12	27	37	18	02	23	30	32	29	27	40	33	43	40	49	47
3. Ability to deal with abstracts			25	26	33	22	11	29	16	22	33	−01	−13	07	28	28	23	20	37	18	40	25	34	34
4. Problem-solving ability				36	42	37	28	27	29	27	29	19	07	21	34	31	32	19	31	24	37	36	46	45
5. Judgment					35	29	33	41	21	41	24	26	13	18	33	31	23	29	34	26	38	28	43	38
6. Foresight						32	20	32	29	32	40	20	12	14	30	29	30	22	29	33	41	33	40	36
7. Decisiveness							19	39	19	29	32	17	09	18	32	32	29	30	37	21	31	43	43	37
8. Accuracy								18	28	45	16	19	11	07	20	20	27	14	23	30	16	16	29	25
9. Identification with the business world									24	35	32	21	13	21	41	42	19	31	44	36	41	40	53	49
10. Identification with the company										29	26	32	27	25	19	26	16	07	27	28	17	26	39	28
11. Dependability											31	23	11	09	31	34	30	25	49	35	33	32	42	34
12. Self-improvement												09	02	10	28	19	28	28	34	26	32	41	35	40
13. Interpersonal relations													53	49	14	34	20	30	15	14	23	23	22	27
14. Sensitivity to others														34	13	25	24	20	08	−05	12	14	15	19
15. Likableness															08	32	06	29	18	18	26	24	27	35
16. Poise																28	26	32	37	12	39	33	41	44
17. Administrative ability																	19	38	47	17	41	38	42	45
18. Written communication																		26	19	08	24	15	25	26
19. Oral communication																			30	17	38	39	36	45
20. Leadership																				39	40	47	52	48
21. Enthusiasm (drive)																					28	43	44	39
22. Persuasiveness																						40	55	52
23. Competitiveness																							49	44
24. Progress																								79
25. Potential																								

Note.—Decimals omitted.

This table was examined to identify promising scales to be used as trial independent variables. Five scale "clusters" were identified, and the scale with the highest median intercorrelation within the cluster was selected. Persuasiveness, Foresight, Dependability, Identification with the Business World, and Likableness were chosen on this basis. Three other scales—Oral Communication, Written Communication, and Identification with the Company—were chosen because of their low correlations with all other scales. Thus eight ratings were used as independent variables.

Partial correlations among the remaining scales were computed with the independent variables held constant. While the median partial r was only .065, 27% of the 105 partials exceeded ±.15. The hypothesis of independence of the partial r's was rejected ($\chi^2 = 351.09$, $df = 136$, $p < .01$).

The eight independent variables correlated .71 and .68 with "Progress" and "Potential" ratings, respectively. But seven of the other dependent variables had partial correlations of .15 or more with these two criteria.

It seemed desirable to effect further reductions in the partial correlations and to increase the correlations with the overall criteria. For these purposes, three additional scales were selected as independent variables: Problem-Solving Ability and Leadership were added because they had partial r's of .20–.24 with the overall criteria; Sensitivity to Others was added because it had several high partials with other independent variables.

Using 11 scales as independent variables, the process described above was repeated for the 14 dependent variables. The median partial correlation was reduced to .035; however, 23% of the partials still exceeded ±.15. Again, the hypothesis of independence among the partial correlations was rejected ($\chi^2 = 213.48$, $df = 91$, $p < .01$).

While the multiple Rs with "Progress" and "Potential" were increased to .74 and .71, respectively, partial correlations of three specific scales with these overall ratings still exceeded .15. Two of these three—Creativity and Enthusiasm (drive)—were chosen as additional independent variables.

When the analysis was repeated using 13 independent variables and 12 dependent variables, the hypothesis of independence among the partial correlations was again rejected ($\chi^2 = 149.87$, $df = 66$, $p < .01$). However, the median partial correlation was reduced to .021, and only 7% of the partials exceeded ±.15. Further, no partial correlation with "potential" ratings was as high as .15 and only one partial with "progress" ratings reached that level. The multiple Rs with these overall criteria were .75 and .72, respectively.

An examination of the entire matrix of partial correlations suggested that the hypothesis of independence may have been rejected solely on the basis of the .58 partial correlation between "progress" and "potential" ratings. To check this possibility, these two ratings were summed and treated as a single overall rating. When the analysis was repeated this time, the hypothesis of independence was accepted ($\chi^2 = 61.26$, $df = 55$, $p > .50$). It was concluded that ratings of "progress" and "potential" were too closely related to serve usefully as separate criteria.

The 13 scales were taken as the best representation of "job performance in general business" which could be made from this study. Their adequacy in accounting for the other 10 specific scales is reviewed in Table 2.

TABLE 2

ADEQUACY OF 13 SPECIFIC RATINGS IN ACCOUNTING FOR 10 OTHER RATINGS OF JOB PERFORMANCE

Dependent variable	Number of partial r's above .15	Partial r[a] with "progress"	Partial r[a] with "potential"
Adaptability	1	17	11
Ability to deal with abstracts	0	00	06
Judgment	0	06	01
Decisiveness	1	10	01
Accuracy	0	02	02
Self-improvement	1	−02	12
Interpersonal relations	0	−10	−05
Poise	1	07	14
Administrative ability	0	01	07
Competitiveness	2	06	−01

[a] Decimals omitted.

TABLE 3

BETA WEIGHTS AND MULTIPLE CORRELATIONS FOR PREDICTING 11 RATINGS OF JOB PERFORMANCE FROM 13 OTHER RATINGS

Dependent variables	Beta weights[a] for predictors[b]													R[c]
	1	2	3	4	5	6	7	8	9	10	11	12	13	
Adaptability		36					15	24						.58
Ability to deal with abstracts								24		−18	19	26		.55
Judgment		22		24				14	18					.55
Decisiveness		20							21		16	15		.52
Accuracy				34		15							17	.50
Self-improvement	22		16				17					20		.50
Interpersonal relations			14		29		13			37				.65
Poise		26						19	18		18		14	.54
Administrative ability		17					16	15		16	28			.60
Competitiveness							24		18		25		25	.62
Overall measure		17	13				16	21	17		14	13	12	.77

[a] Decimals omitted.
[b] 1 = Foresight
2 = Identification with the business world
3 = Identification with the company
4 = Dependability
5 = Likableness
6 = Written communication
7 = Oral communication
8 = Persuasiveness
9 = Problem-solving ability
10 = Sensitivity to others
11 = Leadership
12 = Creativity
13 = Enthusiasm (drive)
[c] Multiple correlation with those predictors whose beta weights were significantly different from zero.

This table shows that the partial correlations among all dependent variables, including the two overall ratings, were low and generally not significantly different from zero.

A step-wise multiple-regression procedure was then followed using the 13 specific scales as predictors of each remaining scale and of the combined "progress" and "potential" scales. Predictors with nonsignificant beta weights were eliminated, one by one, until all betas were significant ($p < .05$). The results are summarized in Table 3.

From Table 3, it can be seen that each of the 13 independent variables carried at least one significant beta weight. Several of them were useful in predicting five or more of the dependent variables.

Eight of the scales made independent contributions to the prediction of the overall criterion. These eight scales are taken as the elements or factors describing success in general business. Table 4 identifies them and describes their relative potency in predicting overall success.

TABLE 4

ELEMENTS OF SUCCESS IN GENERAL BUSINESS ACTIVITIES

Element	Beta weight (b)	r[a]	Variance contribution ($b \times r$)
Persuasiveness	.211	.56	.118
Identification with the business world	.171	.54	.092
Problem-solving ability	.166	.48	.080
Leadership	.137	.53	.073
Oral communication	.159	.43	.068
Creativity	.132	.51	.067
Enthusiasm (drive)	.120	.44	.053
Identification with the company	.127	.35	.044
			$.595 = R^2$; $R = .77$

[a] Single-order correlation with overall criterion.

DISCUSSION

Several limitations should be acknowledged. Results were obtained from graduates of only one university who had from 5 to 10 years postgraduate experience. "Success" was evaluated solely from the employer's point of view. The reliability of the rating scales was unknown, and it was necessary to assume interrater and intercompany comparability. Finally, it is probable that a limited range of "success" was assessed since "colossal fail-

ures" would not be expected to remain in business for 5 or more years. These limitations provide the framework for the following discussion.

Eight factors were found to contribute to the prediction of overall success in business jobs—Persuasiveness, Identification with the Business World, Problem-Solving Ability, Leadership, Oral Communication, Creativity, Enthusiasm (drive), and Identification with the Company. Ratings on these factors correlated .77 with the overall criterion. It seems likely that the unknown reliabilities and limited ranges of both predictor and criterion measures account for most of the remaining criterion variance.

A crucial question is, "Did these *S*s possess these eight characteristics in the same relative amounts at the time of graduation?" Put otherwise, to what degree has postcollege experience differentially modified these personal characteristics? If one assumes that the postcollege experience has not significantly affected the relative standing of the *S*s on these characteristics, then recruiters would be well advised to concentrate their appraisals on such characteristics. The basic assumption underlying this recommendation has not been evaluated in this setting; studies from other settings (Jacob, 1957; Astin, 1963) support it in that they have shown the chief correlates of "output" measures are "input" measures. That is, final measures of the characteristics that these investigators studied reflected initial measures of these same characteristics much more than any special environmental or treatment conditions.

Students and their counselors should find these results to be useful. Perhaps the most immediate implication is the desirability of assessing student potentials in terms of characteristics like persuasiveness, leadership, etc. The lack of objective devices for making such appraisals points both to the need for such devices and the necessity for applying careful clinical analyses in their absence.

Again the question must be raised of the sensitivity of such characteristics to training and educational experiences. Can students judged to be lacking in problem-solving ability, for example, be advised to take certain courses or gain certain experiences to improve these skills? This is an important unanswered question. The "input-output" studies referred to earlier suggest that it would be safer to identify students who already possess the desired characteristics than to provide remedial programs for the "have nots." But it would seem worthwhile to investigate hypotheses such as the following: Courses in mathematics and logic develop problem-solving skills; extracurricular participation develops leadership; speech courses and part-time sales jobs develop persuasiveness; writing and art courses develop creativity.

Job performance was described by the eight elements of job success plus five additional elements: foresight, dependability, likableness, sensitivity to others, and written communication. College graduates employed in general business activities can be differentiated by these five performance attributes, but this differentiation does not contribute to an assessment of their job success when the other eight performances have been taken into account. When an employee is described as "dependable," "likable," etc., some aspect of his performance is being described which is not redundant with other descriptions ("persuasive," "creative," etc.). However, if the employee were rated on the eight items identified as elements of job success, further ratings of his dependability, likableness, etc., would not contribute to a judgment of overall success on the job.

The results provide some empirical data on the criterion problem in one occupational setting. The complexity of the "success" criterion, and its identified dimensions, were forcefully illustrated. Since employee selection and recruitment and student guidance programs all require the ability to forecast "success," it is essential that further efforts be made to define this concept more explicitly in a variety of occupational settings.

REFERENCES

ANDERSON, T. W. *Introduction to multivariate statistical analysis.* New York: John Wiley, 1958.

ASTIN, A. W. Differential college effects on the motivation of talented students to obtain the Ph.D. *Journal of Educational Psychology*, 1963, **54**, 63–71.

BECHTOLDT, H. P. Factor analysis and the investigation of hypotheses. *Perceptual & Motor Skills*, 1962, 14, 319–342.

BECHTOLDT, H. P., BENTON, A. L., & FOGEL, M. L. An application of factor analysis in neuropsychology. *Psychological Record*, 1962, 12, 147–156.

DICKENSON, C. What employers look for in the college graduate. *Personnel and Guidance Journal*, 1955, 33, 460–464.

JACOB, P. E. *Changing values in college*. New York: Harper, 1957.

JOHNSON, J. What we look for in the college graduate. *Journal of College Placement*, 1959, 19, 30–34.

PALLETT, J. E. Definition and prediction of success in the business world. Unpublished doctoral dissertation, University of Iowa, 1965.

SUPER, D. E. *The psychology of careers*. New York: Harper, 1957.

SUPER, D. E., & CRITES, J. O. *Appraising vocational fitness by means of psychological tests*. New York: Harper & Row, 1962.

THORNDIKE, R. L., & HAGEN, E. *Measurement and evaluation in psychology and education*. New York: John Wiley, 1961.

(Received May 2, 1966)

(Continued from page 151)

Use of the Weighted Application Blank in Selecting Unskilled Employees: Richard D. Scott and Richard W. Johnson*: Department of Psychology, University of Massachusetts, Amherst, Massachusetts.

The Multitrait-Multirater Approach to Measuring Managerial Job Performance: Edward E. Lawler, III*: Department of Industrial Administration, Yale University, 2 Hillhouse Avenue, New Haven, Connecticut 06520.

Effects of Inequity Produced by Underpayment on Work Output, Work Quality, and Attitudes toward the Work: Edward E. Lawler, III* and Paul W. O'Gara: Department of Industrial Administration, Yale University, New Haven, Connecticut 06520.

Effects of Monetary Rewards and Punishments on Vigilance Performance: Russell L. Smith*, Luigi F. Lucaccini and Murray H. Epstein: Biotechnology Laboratory, Room 3116, Engineering, UCLA, 405 Hilgard Avenue, Los Angeles, California 90024.

A Multidimensional Scaling Analysis of the Job of Civil Defense Director: Robert J. Smith* and Arthur I. Siegel: Education Center, 36th Combat Support Group, United States Air Force, APO New York 09403.

Comparison of Performance of Sales Training Graduates and Nongraduates: Ward A. Harris* and Norman L. Vincent: Agency Research Department, State Farm Mutual Automobile Insurance Company, 112 East Washington Street, Bloomington, Illinois 61701.

---

* Asterisk indicates author for whom the address is supplied.

Journal of Applied Psychology
1967, Vol. 51, No. 2, 181-186

EFFECTS OF APTITUDE-SCORE ADJUSTMENTS BY AGE CURVES ON PREDICTION OF JOB PERFORMANCE

ROBERT C. DROEGE

United States Employment Service, Washington, D. C.

This study was conducted by the United States Employment Service to determine the relative validity of unadjusted and age-adjusted GATB aptitude scores for predicting occupational success. 11 longitudinal occupational validation studies, conducted on samples varying in size from 56 to 124 cases, were selected for the analysis. For each sample the validities of unadjusted and age-adjusted aptitude scores for predicting occupational success were compared for the 9 GATB aptitude measures. Differences between validities tended to be small.

The average age of the nation's labor force continues to increase, and growing numbers of older workers are being involuntarily retired or technologically displaced and compelled to look for other jobs late in their working lives. Many of those who need help in finding suitable new employment will not be able to obtain jobs closely related to those they had previously. To what extent can aptitude testing help determine what other occupations these older persons can learn easily and perform successfully?

Since most of the research on development of aptitude batteries for personnel selection and counseling has been carried out with samples of younger workers, a question arises as to the applicability of these test batteries for older persons. A full investigation of this question requires study of (*a*) the relationship between aptitudes and age, (*b*) the relationship between job performance and age, and (*c*) the interrelationships of aptitudes, job performance, and age.

This article is concerned with research on one aspect of interrelationships of aptitudes, job performance, and age. Specifically, the research is on validity of age adjustments for aptitudes in prediction of job performance. The age adjustments referred to are those typically used to develop age norms, so that the average adjusted test score is the same for all age groups. The question of the desirability of making such adjustments to develop age norms for intelligence and aptitude tests is an old one which has not been resolved. Ultimately, the basis for deciding between unadjusted and age-adjusted aptitude scores lies in the relative validity of the two sets of scores in prediction of performance for appropriate occupational samples.

Experimental Design Requirements

A study of the relative validity of unadjusted and age-adjusted aptitude scores requires a longitudinal rather than a cross-sectional design for the following reasons:

1. Maintenance of ability once acquired is not the same as acquisition of a new ability. The typical situation in cross-sectional studies is that the employed workers have been trained and have acquired the basic skills required on the job. Job skills of these experienced workers are constantly being reinforced with experience so that, for some occupations, performance may not change to any extent with age. However, ability to *learn* the job may change considerably with age, and learning ability may be a factor of great importance to the employer.

2. In cross-sectional studies it is often difficult to obtain comparable measures of job performance for individuals in the various age intervals. For example, the older, more experienced workers may tend to occupy the most favorable work positions so that their performance cannot be compared directly to the performance of younger workers. Similarly, supervisory ratings may be biased in favor of the more experienced (and older) workers.

TABLE 1

INFORMATION ON STUDIES INCLUDED IN THE ANALYSIS

Occupation	DOT code	N	Type of criterion
Assembler, dry-cell and battery	727.887	94	Supervisory ratings
Automobile-body repairman, metal	807.381	56	Instructor's ratings
Clerk-stenographer	202.388	118	Instructor's ratings
Conveyor operator	921.883	64	Supervisory ratings
Laborer, ammunition assembly II	737.887	73	Supervisory ratings
Mail sorter	231.688	80	Work sample
Maintenance man, building	899.381	86	Instructor's ratings
Mill inspector	619.381	70	Supervisory ratings
Psychiatric aid	355.878	73	Instructor's ratings
Hand sewer, shoes	788.884	64	Instructor's ratings
Tool-and-die maker	601.280	124	Supervisory ratings

3. As pointed out in a report on a cross-sectional study of age and job performance of Federal Mail Sorters (Walker, 1964):

> The composition of the youngest age group is not strictly comparable to that of succeeding age groups. The youngest group includes the ambitious and the superior as well as the less ambitious and inept. Through the years, the characters of the age groups change as the substandard are discharged and some of the superior are promoted. The oldest age groups of sorters, therefore, contain more of the acceptable workers with longer experience but fewer of the superior workers since these were selected for promotion or transfer to more attractive work [p. 297].

This is the typical situation in cross-sectional studies, with little or no basis for an assumption of comparability of the younger and older groups.

The ideal design for investigating the applicability of adjustments in aptitude scores for age should have the following characteristics:

1. The design should be longitudinal, that is, individuals in the sample should be tested with the aptitude tests before they are hired or selected for training, but test results should not be a factor in the selection.

2. Only applicants without previous experience or training in the occupation studied should be included in the sample. This is an important requirement because aptitude tests are designed for use in predicting ability to acquire *new* skills.

3. The sample should cover a wide age range to permit meaningful comparison of unadjusted and age-adjusted sets of scores.

4. The criterion should be a measure of training success or early job success.

PROCEDURE

The General Aptitude Test Battery (GATB) was used as the experimental test battery for this research. The GATB consists of nine aptitudes measured by 12 tests as follows:

Aptitude	*Tests*
G—Intelligence	Three-Dimensional Space
	Arithmetic Reason
	Vocabulary
V—Verbal Aptitude	Vocabulary
N—Numerical Aptitude	Arithmetic Reason
	Computation
S—Spatial Aptitude	Three-Dimensional Space
P—Form Perception	Form Matching
	Tool Matching
Q—Clerical Perception	Name Comparison
K—Motor Coordination	Mark Making
F—Finger Dexterity	Assemble
	Disassemble
M—Manual Dexterity	Place
	Turn

Studies on the relationship between GATB aptitude scores and age for adults (Droege, Crambert, & Henkin, 1963) show that all aptitudes except Verbal Aptitude (V) decline with age and that there is variation among the aptitudes with regard to age of onset of decline. In these studies age curves were constructed showing the relationship between average scores on each aptitude and age over an age range from 17 to 72. Availability of reasonably good age-curve data for GATB aptitudes made possible the development of age norms for the aptitudes. That is, obtained scores could be adjusted so that the average adjusted aptitude score was the same throughout the age range. Since there was a marked decline in obtained scores with age on most of the aptitudes, the corresponding adjustments were substantial for older individuals.

The United States Employment Service (USES) has a continuing program of aptitude-test validation studies on specific occupations, and the entire GATB is used as the experimental test battery in each study. Eleven of these studies were selected for analyses to determine relative validity of unadjusted and age-adjusted aptitude scores. Basic information on these studies is shown in Table 1. The following characteristics of these studies should be noted:

1. All were conducted using a longitudinal design; that is, the individuals in the samples were tested prior to entry on the job or into the training course, but the aptitude scores were not used in selection.

2. Six of the samples consisted of employed workers in specific occupations; five consisted of trainees for specific occupations.

TABLE 2

MEAN, STANDARD DEVIATION, RANGE, AND PRODUCT-MOMENT CORRELATION WITH THE CRITERION (r) FOR YEARS OF AGE AND YEARS OF EDUCATION

Occupation	Years of age				Years of education			
	M	SD	Range	r	M	SD	Range	r
Assembler, dry-cell and battery	29.0	9.1	18–51	.03	9.6	1.9	6–14	.10
Automobile-body repairman, metal	28.6	7.5	18–51	−.03	10.4	1.7	6–14	−.02
Clerk-stenographer	26.0	9.5	17–58	−.34**	12.0	.7	9–14	−.09
Conveyor operator	36.8	7.7	23–54	.14	10.7	2.5	1–16	.10
Laborer, ammunition assembly II	35.3	6.9	21–51	−.36**	11.6	.9	8–12	.04
Mail sorter	33.1	9.4	18–53	.16	12.7	1.7	9–16	.13
Maintenance man, building	41.9	9.6	19–59	.07	9.8	1.6	6–12	.01
Mill inspector	29.2	8.6	18–52	−.15	12.1	2.4	6–17	.25*
Psychiatric aid	25.3	7.5	17–52	.16	12.1	.7	10–14	.04
Hand sewer, shoes	30.5	13.3	16–60	.12	9.0	2.0	6–14	.04
Tool-and-die maker	23.5	4.0	19–37	−.16	12.1	.8	10–16	.18*

* $p < .05$.
** $p < .01$.

3. The last three digits of the DOT (3rd edition) code reflect the complexity of the occupation (United States Department of Labor, Bureau of Employment Security, 1965). A wide range of complexity is represented in the 11 samples.

4. The criteria consisted of ratings for all except the Mail-Sorter sample for which a work-sample criterion was obtained. For some of the samples, pooled ratings, representing the composite judgment of two or more raters, were obtained. For other samples, only one individual did the rating. Reliability data on the criteria were not available.

Table 2 shows data on age and education for each of the samples. Note that there is a relatively wide range of age for each sample. This was an important consideration in the selection of the studies for the present investigation. Two of the studies have significant negative correlations between age and the criterion and two others have significant positive correlations between education and the criterion.

RESULTS

The mean unadjusted GATB aptitude score is 100 and the standard deviation is 20 for the GATB general working-population sample (United States Department of Labor, Bureau of Employment Security, 1962). Means and standard deviations of unadjusted and age-adjusted aptitude scores for the 11 samples are shown in Table 3. The age adjustments were made from the age curves obtained from the four-state sample in the study on the relationship between GATB aptitude scores and age for adults (Droege, Crambert, & Henkin, 1963). For each aptitude, the age group with the highest average score was identified. This average score was then used as the base point for making score adjustments for individuals in other age groups.

The data in Table 3 show that there is variation among samples with regard to the extent to which age adjustments affect means and standard deviations of aptitude scores. There is similar variation among the aptitudes for the separate samples. Notice that the age adjustments always result in decreases in mean scores.

Table 4 shows the correlation between the criterion of performance on the job or in training and the unadjusted and age-adjusted aptitude scores. Many of the validity coefficients are statistically significant for both unadjusted and age-adjusted aptitude scores. Table 5 shows the differences in validity coefficients of the age-adjusted and unadjusted aptitude scores. Small differences are significant in some instances because of the high correlations between unadjusted and age-adjusted scores, which have an important effect on the significance level (Guilford, 1965).

Table 6 shows the distribution of the validity differences for all 9 aptitudes and 11 samples. The median difference is .00.

TABLE 3

MEANS AND STANDARD DEVIATIONS OF UNADJUSTED AND ADJUSTED GATB APTITUDE SCORES FOR THE 11 SAMPLES

Occupation	Aptitudes of the General Aptitude Test Battery								
	G	V	N	S	P	Q	K	F	M
Assembler, dry-cell and battery									
M	90.7	90.3	90.7	94.4	97.6	97.5	96.4	104.1	100.0
	92.5	*94.2*	*92.3*	*99.5*	*103.8*	*101.6*	*99.8*	*104.2*	*103.5*
SD	13.4	12.5	15.7	16.5	17.8	15.7	15.8	19.2	18.4
	13.2	*13.1*	*15.4*	*15.4*	*15.3*	*14.8*	*16.0*	*19.4*	*18.8*
Automobile-body repairman, metal									
M	96.3	93.3	89.7	110.2	96.9	95.7	88.7	104.6	103.6
	97.5	*96.9*	*91.2*	*114.6*	*102.4*	*98.8*	*91.3*	*106.9*	*106.3*
SD	15.3	15.0	15.0	17.2	15.9	15.5	20.2	19.9	19.9
	15.2	*14.0*	*14.9*	*17.1*	*16.1*	*15.3*	*20.7*	*19.8*	*19.4*
Clerk-stenographer									
M	97.4	97.0	100.4	98.4	103.8	105.6	112.2	100.1	103.1
	98.9	*101.6*	*101.7*	*101.6*	*107.9*	*108.5*	*114.7*	*102.4*	*106.1*
SD	14.9	12.3	14.9	17.9	20.0	14.2	16.4	19.4	20.4
	15.1	*12.4*	*14.9*	*18.1*	*18.2*	*13.6*	*15.8*	*18.3*	*19.7*
Conveyor operator									
M	92.1	87.2	90.9	97.4	87.4	74.1	72.6	94.1	94.3
	94.2	*89.2*	*93.0*	*105.9*	*99.7*	*81.1*	*79.0*	*100.9*	*101.1*
SD	15.7	11.9	16.9	18.1	16.6	14.2	17.5	20.4	18.6
	16.3	*12.2*	*16.8*	*17.6*	*15.5*	*14.2*	*16.9*	*19.3*	*17.9*
Laborer, ammunition assembly II									
M	93.1	92.3	89.8	98.4	90.8	85.7	89.8	113.4	117.6
	95.1	*94.6*	*91.9*	*105.8*	*101.5*	*91.7*	*95.3*	*118.7*	*123.2*
SD	15.5	16.5	18.8	17.0	16.4	16.5	19.8	16.2	17.9
	15.7	*16.6*	*18.8*	*16.5*	*15.7*	*15.8*	*19.3*	*14.7*	*17.3*
Mail sorter									
M	97.2	100.1	98.8	97.4	102.2	105.9	107.0	98.8	107.8
	98.9	*102.9*	*100.6*	*99.0*	*111.7*	*111.3*	*111.7*	*104.9*	*113.1*
SD	14.7	15.1	15.7	17.4	16.6	14.9	16.0	17.8	19.0
	14.9	*14.7*	*15.6*	*17.6*	*16.2*	*14.8*	*16.6*	*18.4*	*19.5*
Maintenance man, building									
M	88.8	91.2	85.4	91.6	82.1	88.7	84.9	75.4	88.5
	91.9	*93.1*	*88.2*	*102.1*	*98.6*	*98.3*	*94.0*	*87.1*	*100.4*
SD	15.3	14.5	17.0	16.2	17.9	12.8	21.8	19.5	23.6
	15.7	*14.4*	*17.3*	*16.4*	*18.0*	*13.5*	*21.0*	*19.7*	*23.8*
Mill inspector									
M	103.5	92.3	101.7	109.0	96.8	85.2	83.2	98.6	99.7
	105.1	*95.9*	*103.1*	*114.1*	*103.2*	*88.8*	*86.4*	*101.8*	*103.1*
SD	20.0	16.9	18.6	19.1	15.5	14.4	18.2	19.9	19.3
	19.8	*16.8*	*18.6*	*18.2*	*14.0*	*14.1*	*17.6*	*18.4*	*18.3*
Psychiatric aid									
M	106.0	104.9	102.0	108.5	111.5	106.5	108.2	104.2	112.0
	107.1	*109.3*	*103.0*	*111.6*	*115.1*	*108.8*	*109.9*	*105.8*	*114.1*
SD	13.4	13.5	12.9	16.2	16.5	12.5	14.5	20.1	27.4
	13.6	*13.5*	*13.1*	*16.5*	*16.9*	*12.6*	*14.2*	*19.6*	*26.5*
Hand sewer, shoes									
M	90.1	86.7	86.4	100.0	94.7	93.1	95.4	98.5	113.8
	92.2	*89.5*	*88.3*	*105.8*	*103.0*	*98.5*	*100.3*	*104.1*	*120.0*
SD	16.3	13.9	18.8	19.0	18.7	13.5	18.6	25.1	25.3
	16.4	*16.1*	*19.1*	*19.4*	*19.3*	*14.2*	*19.8*	*23.9*	*24.6*
Tool-and-die maker									
M	111.7	98.8	107.8	118.5	111.7	96.5	102.5	104.5	122.4
	112.6	*103.6*	*108.6*	*120.6*	*113.6*	*97.8*	*103.3*	*105.0*	*123.4*
SD	14.8	13.7	14.7	16.3	15.3	15.6	16.6	18.0	17.4
	15.0	*13.9*	*14.9*	*16.4*	*15.3*	*15.7*	*16.7*	*18.1*	*17.4*

Note.—*M*s and *SD*s of *adjusted* GATB aptitude scores are in italics.

TABLE 4

CORRELATIONS BETWEEN THE CRITERION AND THE UNADJUSTED AND ADJUSTED GATB APTITUDE SCORES FOR THE 11 SAMPLES

Occupation	Aptitudes of the General Aptitude Test Battery								
	G	V	N	S	P	Q	K	F	M
Assembler, dry-cell and battery	.11	.20	.09	.11	.27**	.16	.45**	.48**	.50**
	.11	*.15*	*.09*	*.14*	*.29***	*.16*	*.43***	*.47***	*.46***
Automobile-body repairman, metal	.09	−.03	.10	.13	.35**	.27*	.19	.17	.29*
	.09	*.04*	*.12*	*.14*	*.34**	*.26**	*.18*	*.14*	*.26*
Clerk-stenographer	.50**	.47**	.49**	.40**	.46**	.48**	.26**	.34**	.25**
	*.48***	*.50***	*.48***	*.33***	*.39***	*.42***	*.19**	*.26***	*.17**
Conveyor operator	.26*	.10	.16	.45**	.41**	.14	.32**	.59**	.55**
	*.26**	*.07*	*.18*	*.48***	*.48***	*.17*	*.36***	*.66***	*.60***
Laborer, ammunition assembly II	.37**	.04	.55**	.33**	.38**	.42**	.38**	.68**	.55**
	*.35***	*.05*	*.55***	*.28*	*.28**	*.35***	*.33***	*.61***	*.46***
Mail sorter	.37**	.48**	.35**	.10	.34**	.32**	.38**	.21	.12
	*.38***	*.48***	*.37***	*.14*	*.38***	*.35***	*.40***	*.23**	*.14*
Maintenance man, building	.33**	.35**	.37**	.19	.35**	.24*	.04	.27*	.18
	*.32***	*.34***	*.37***	*.21**	*.39***	*.26**	*.05*	*.28***	*.18*
Mill inspector	.44**	.30*	.42**	.42**	.31**	.27*	.16	.23	.29*
	*.45***	*.32***	*.42***	*.43***	*.28**	*.25**	*.13*	*.20*	*.25**
Psychiatric aid	.43**	.56**	.30**	.08	.01	.22	.24*	.13	−.03
	*.44***	*.54***	*.32***	*.13*	*.05*	*.25**	*.25**	*.13*	*−.04*
Hand sewer, shoes	.25*	.08	.20	.35**	.15	.20	.19	.36**	.33**
	.23	*.09*	*.20*	*.37***	*.16*	*.18*	*.17*	*.36***	*.30**
Tool-and-die maker	.35**	.09	.37**	.45**	.48**	.28**	.17	.23*	.30**
	*.35***	*.10*	*.37***	*.43***	*.46***	*.27***	*.17*	*.22**	*.30***

Note.—Correlations between the criterion and the *adjusted* GATB aptitude scores are in italics.
* $p < .05$.
** $p < .01$.

TABLE 5

DIFFERENCES IN VALIDITY COEFFICIENTS OF AGE-ADJUSTED AND UNADJUSTED APTITUDE SCORES FOR THE 11 SAMPLES

Occupation	Aptitudes of the General Aptitude Test Battery								
	G	V	N	S	P	Q	K	F	M
Assembler, dry-cell and battery	.00	.05	.00	−.03	−.02**	.00	.02**	.01	.04
Automobile-body repairman, metal	.00	−.07	−.02	−.01	.01	.01	.01	.03	.03
Clerk-stenographer	.02	−.03**	.01	.07**	.08**	.06**	.07**	.08**	.08**
Conveyor operator	.00	.03	−.02	−.03	−.07	−.03	−.04	−.07*	−.05
Laborer, ammunition assembly II	.02*	−.01	.00	.05*	.10*	.07**	.05**	.07	.09**
Mail sorter	−.01	.00	−.02*	−.04	−.04	−.03	−.02	−.02	−.02
Maintenance man, building	.01	.01	.00	−.02	−.04	−.02	−.01	−.01	.00
Mill inspector	−.01	−.02	.00	−.01	.03	.02	.03	.03	.04
Psychiatric aid	−.01	.02	−.02	−.05	−.04	−.03	−.01	.00	.01
Hand sewer, shoes	.02	−.01	.00	−.02	−.01	.02	.02	.00	.03
Tool-and-die maker	.00	−.01	.00	.02	.02	.01	.00	.01	.00

Note.—Positive differences indicate validity of unadjusted is higher; negative differences indicate validity of age adjusted higher.
* $p < .05$.
** $p < .01$.

TABLE 6

DISTRIBUTION OF DIFFERENCES IN VALIDITY COEFFICIENTS OF AGE-ADJUSTED AND UNADJUSTED APTITUDE SCORES

Differences in validity (unadjusted − adjusted) Interval	Midpoint	Frequency
.08 to .10	.09	5
.05 to .07	.06	8
.02 to .04	.03	19
−.01 to .01	.00	39
−.02 to −.04	−.03	23
−.05 to −.07	−.06	5

DISCUSSION

The results indicate that, in some instances, aptitude scores adjusted for age have validities that are significantly different from validities of unadjusted scores. The sign and direction of the difference are dependent on the occupation, the age range represented in the sample, and the particular aptitude. These factors are discussed below:

1. With reference to the occupations included in this investigation, age adjustments in the aptitude scores resulted in substantially higher validity for only one occupation (Conveyor Operator), but such adjustments resulted in substantially lower validities for two occupations (Clerk-Stenographer and Laborer, Amunition Assembly II). For the remaining eight occupations the differences between age-adjusted and unadjusted score validities were quite small.

2. An important consideration in the selection of the particular 11 samples for this investigation was the age range represented in the sample. An attempt was made to include samples with high variability in years of age so that differences in unadjusted and age-adjusted sets of scores would be maximized. Obviously, in studies based on samples which do not have a wide age range, substantial differences between validities of unadjusted and age-adjusted scores are not so likely to occur.

3. The aptitudes for which smallest differences in validity of unadjusted and age-adjusted scores were Intelligence, Verbal Aptitude, and Numerical Aptitude. For these aptitudes validity appears to be largely unaffected when aptitude scores are adjusted for age. The primary reason for this appears to be the fact that these three aptitudes show little or no decline with age, whereas the other aptitudes all show relatively sharp decline in average score with age after the age of onset of decline.

The present investigation was limited to comparisons of validity coefficients of unadjusted and age-adjusted aptitude scores. The USES plans an additional study focusing on the individuals who fail established aptitude cutting scores for specific occupations when unadjusted scores are used but pass the same cutting scores when age-adjusted scores are used. Analyses will be done to determine whether these workers who shift from a fail to a pass category are more like successful or unsuccessful workers.

REFERENCES

DROEGE, R. C., CRAMBERT, A. C., & HENKIN, J. B. Relationship between GATB aptitude scores and age for adults. *Personnel and Guidance Journal*, 1963, 42, 502–508.

GUILFORD, J. P. *Fundamental statistics in psychology and education*. (4th ed.) New York: McGraw-Hill, 1965.

UNITED STATES DEPARTMENT OF LABOR, BUREAU OF EMPLOYMENT SECURITY. *Guide to the use of the General Aptitude Test Battery*, Section III: Development. Washington: United States Government Printing Office, 1962.

UNITED STATES DEPARTMENT OF LABOR, BUREAU OF EMPLOYMENT SECURITY. *Dictionary of occupational titles*. (3rd ed.) Washington: United States Government Printing Office, 1965.

WALKER, J. F. The job performance of federal mail sorters by age. *Monthly Labor Review*, March 1964, 296–300.

(Received May 20, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 187-190

SONAR TARGET DETECTION AS A DECISION PROCESS [1]

W. P. COLQUHOUN

Medical Research Council, Applied Psychology Research Unit, Cambridge, England

In a vigilance task which simulated sonar target detection, 12 Ss were required to report the occurrence of 200 msec pulses in a 900 cps tone, which was continuously modulated in amplitude by a low-bandwidth noise source. In 3 45-min sessions Ss reported signals only when quite certain of their identification (Sure procedure); in a further 3 sessions they reported any signal-like sound (Unsure procedure). A substantially higher percentage of both "weak" and "strong" signals was detected with the Unsure procedure. False report rate was also higher with the Unsure procedure, but analysis showed that this reflected a change in decision criterion rather than in discrimination efficiency. Within-session decrement was slight, and unrelated either to signal strength or reporting procedure. The results support the contention that decision processes play a major role in determining performance at this kind of task

In recent years considerable interest has been shown in the applicability of the theory of signal detectability (Swets, Tanner, & Birdsall, 1961) to behavior in tasks which require the operator to keep watch for weak and relatively rare signals, such as radar and sonar targets, in a vigilance setting (Broadbent & Gregory, 1963; Jerison & Pickett, 1963; Mackworth & Taylor, 1963). This theory considers detection as essentially a decision process based on a statistical criterion; the criterion may be altered by a number of factors, including the operator's risk-taking "set." For example, Evans (1965) showed that, when instructed to adopt a risky reporting set, subjects (*Ss*) made earlier detections and gave more false reports, than when instructed to adopt a cautious reporting set.

In the Evans study a video display was employed to simulate inbound radar tracks. A certain degree of temporal uncertainty was introduced by varying the range at which the target first became visible, but the task was not one of vigilance in the usually accepted sense, since each target was presented on a separate trial in the presence of an examiner who gave full knowledge of results. The present study was also concerned with the effect of reporting set on target detection, but here performance was observed in a typical vigilance situation, that is, in a prolonged session with no knowledge of results, and the signals were monitored auditorily, rather than on a video display, in order to simulate certain sonar operations where this mode of presentation is employed.

Method

Apparatus

The sonar target simulator used in this experiment was an electronic device with a variable signal-to-noise amplitude (S/N) ratio capability. It generated a continuous 900-cycle-per-second tone, amplitude modulated by a low bandwidth (0–10 cycles-per-second) random noise source. The target "signals" consisted of pulses produced by removing the noise modulation and simultaneously raising the tone amplitude for approximately 200 milliseconds. Four pulse amplitudes, the values of which resulted in S/N ratios of 9.7, 10.8, 11.8, and 12.6 decibels ("noise" amplitude being taken as the root mean square of the modulated tone), were employed to give a range of signal strengths from "weak" to "strong." Pulse rise-time was set at a value which obviated sharp transients. The simulator output was presented over earphones to *S*, who was instructed to report detection of a signal by closing a contact switch mounted in front of him. Responses made later than 3 seconds after a signal had occurred were counted as false reports.

Subjects and Procedure

The *Ss* were 12 enlisted men of the Royal Navy, unfamiliar with this form of monitoring. Each *S* was given several hours training, following which he

[1] This research was supported by the Royal Naval Personnel Research Committee of the British Medical Research Council. The cooperation of the Royal Naval Unit, Cambridge, is gratefully acknowledged.

performed the task for 8 sessions, each of 45 minutes' duration, at the rate of 2 sessions per day. Sessions 2, 3, 4, and 6, 7, 8 were regular "test" sessions: Sessions 1 and 5 were "instructional" sessions during which *S* was taught to adopt one of the following procedures for reporting signals:

"Sure" procedure: S was required to report a signal only when he was quite certain that he had indentified a sound as such.

"Unsure" procedure: S was required to report any sound which he considered might be a signal.

Six *Ss* were taught the Sure procedure on Session 1 and the Unsure procedure on Session 5; the other 6 *Ss* were taught in the reverse order. In each of the sessions 36 signals were presented. Intersignal intervals were randomly determined within the constraint that 3 signals of each strength occurred in each successive 15-minute period of the session. The order in which the different signal strengths were represented was randomized within each period.

Results

The proportion of signals detected with the Unsure procedure was substantially greater than the proportion detected with the Sure procedure at each of the four signal strengths. Analysis of variance of the overall detection scores showed that the main effects of procedure and of signal strength were both statistically significant, $F\ (1/10) = 30.5$, $p < .001$; and $F\ (3/66) = 74.1$, $p < .001$, respectively. However, there was no significant degree of interaction between these effects. The relation between detection score and signal strength was apparently linear with both procedures (Figure 1). No significant effect of practice was observed, but the

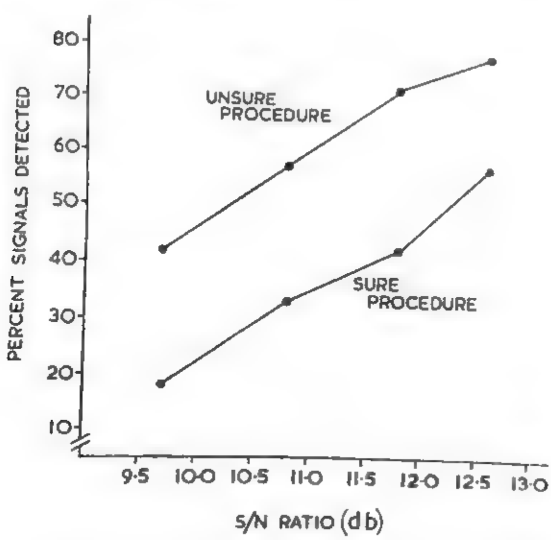


Fig. 1. Percentage of signals detected at four levels of signal strength, with two reporting procedures.

TABLE 1

Overall Performance Scores with Sure and Unsure Procedures

Performance measure	Sure	Unsure
Percentage of signals detected	37.6	61.8
Median no. false reports	16.0	56.0
Median d'	2.2	2.3
Median β	20.0	5.6

overall mean detection score for the six *Ss* who performed first with the Sure procedure was significantly lower than the score for the six *Ss* who performed first with the Unsure procedure, $F\ (1/10) = 6.6$, $p < .05$. This may have been caused by a "carry-over" effect of the first procedure experienced upon performance with the second, that is, to a partial failure of the second instructional session to overcome the pattern of behavior set by the first. However, all 12 *Ss* returned a higher detection score with the Unsure procedure than with the Sure procedure, regardless of the order in which they were tested.

The median false report rate with the Unsure procedure was 3.5 times as high as the rate with the Sure procedure (Table 1). Of the 12 *Ss*, 11 made more false reports in the former condition; the difference between procedures was statistically significant ($p < .01$, Wilcoxon matched-pairs signed-ranks test).

The substantial increase in detection rate with the Unsure procedure cannot reasonably be accounted for on the supposition that the accompanying rise in false reports reflected a higher incidence of random "guessing." Even with the relatively high response rate elicited by this procedure the chance probability of a report being made within 3 seconds of a signal was only 0.04 (this figure was computed on the assumption that no more than one report could be made in any 3-second period, which was empirically correct).

Given the assumption of a maximum response rate of one per 3 seconds, it can be shown that the probability of making a false report was 0.007 with the Sure procedure and 0.029 with the Unsure procedure. The difference of only 0.022 resulting from the alteration in procedure should be compared with the much greater change in the probability of

TABLE 2

PERCENTAGE OF SIGNALS DETECTED IN SUCCESSIVE 15-MINUTE PERIODS OF THE TEST SESSIONS

Reporting procedure	Signal strength	Within-session period		
		First	Second	Third
Sure	1 (weak)	20.4	18.5	13.9
	2	38.9	31.5	29.6
	3	46.3	29.6	50.9
	4 (strong)	59.3	53.7	58.3
Unsure	1 (weak)	44.4	42.6	38.9
	2	60.2	48.1	63.9
	3	78.7	66.7	67.6
	4 (strong)	84.3	72.2	74.1

detection (Table 1); here the difference was 0.242, which is over 10 times as large.

Whereas this relation between detections and false reports is quite inconsistent with the use of a correction for the latter based on random guessing, it conforms quite well to the predictions of the theory of signal detectability, in which each report is considered as a decision based on a statistical criterion (Swets et al., 1961). Analysis of the results in terms of this theory (ignoring signal strength and assuming a "decision interval" of 3 seconds with both procedures[2]) showed that β (level of decision criterion) was significantly greater with the Sure procedure than with the Unsure procedure ($p < .01$, Wilcoxon test), reflecting the higher degree of "caution" exercised in reporting in the former case (Table 1). The analysis also showed that d' (signal detectability) was not significantly different with the two procedures, demonstrating that the actual efficiency with which signals were *discriminated* was unaffected by the change in reporting procedure.

Within-Session Changes in Detection Rate

Detection rates in successive 15-minute periods of the test sessions were examined for evidence, first, that any marked within-session decrement occurred, and, second, that the extent of such decrement was related either to the reporting procedure or to the signal strength (theories of vigilance that attempt to account for decrement as an increase in sensory threshold during the test session would predict greater decrement at the lower values of the latter). These detection rates are given in Table 2, from which the following conclusions may be drawn: first, some decrement did occur during test sessions, but its extent was relatively small; second, the decrement did not appear to be affected by the change in reporting procedure; and third, it was not related in any systematic way to the signal strength.

Pooled within-session performance changes are shown in Table 3. The greatest decrement in detection score occurred between the first and second test periods; the false report rate also declined during this time. Both changes in score were statistically significant ($p < .01$, Wilcoxon tests). Whereas d' was virtually identical in the two periods, β rose considerably; however, the increase in β just failed to reach statistical significance at the 5% level of confidence (Wilcoxon test).

DISCUSSION

It appears that a substantial increase in detection rate in vigilance tasks such as sonar target detection can be achieved by encouraging *Ss* to report signals even when they are uncertain as to their identification. The results obtained show that the improvement in performance applies equally to weak and strong signals. Analysis in terms of the theory of signal detectability indicates that the rise in false report rate that follows the adoption of a "risky" reporting procedure does not

TABLE 3

WITHIN-SESSION TRENDS IN VARIOUS PERFORMANCE MEASURES, POOLED OVER ALL SESSIONS AND SIGNAL STRENGTHS

Performance measure	Within-session period		
	First	Second	Third
Percentage of signals detected	54.1	45.4	49.7
Median no. false reports	28.5	20.0	23.0
Median d'	2.2	2.1	2.2
Median β	7.6	11.0	10.0

[2] Although the absolute values of the statistics computed in applications of the theory of signal detectability depend on the assumed decision-interval, it has been shown that estimates of *changes* in these statistics are only slightly affected by considerable error in its assessment, even to the extent of a 30-fold discrepancy (Mackworth, 1965)

represent a change in the actual efficiency with which signals are discriminated, but, rather, a shift in the criterion used to make decisions about responding. This demonstration that performance can readily be manipulated by variation in instructions emphasizes the importance of careful control over the training given to *Ss* in experiments on vigilance, and supports previous work in showing that attempts to construct theories of behavior in these tasks must take into account the decision processes involved. In practical situations, the question of whether an increased rate of detection of real targets is acceptable when accompanied by a higher probability of an incorrect identification is one that must be determined on the basis of the relative values and costs involved.

The fact that there was a positive relationship between signal strength and probability of detection is perhaps not surprising, in view of the similar findings of earlier studies on this point (reviewed by Jerison & Pickett, 1963); however, it is worthy of note that in these studies the effect of signal strength was typically investigated by changing its value in separate trials, rather than, as here, by comparing detection rates for different levels in the same session. Also of interest in the present result is the apparently close approximation of the relationship to a simple linear function over the range of signal strengths covered.

The finding that within-session decrement was relatively slight is consistent with the hypothesis advanced elsewhere (Colquhoun & Baddeley, 1964) that much of the decrement usually shown in vigilance tasks of this kind is due to the practice of testing *Ss* once only, without providing adequate prior experience of the average probability of signal occurrence. When, as in the present case, *Ss* are tested repeatedly, with an invariant signal probability, this major cause of decrement is obviated. However, a certain "residual" decrement may, as here, still occur.

The true cause of this residual decrement remains to be identified. The fact that in this experiment its magnitude was unrelated to signal strength would seem to disprove, at least, the argument that it reflects a rise in sensory threshold due to some sort of "fatigue." The indications in the present results were that the decrement is produced by a shift in decision criterion rather than by a change in signal detectability. Whereas such a conclusion would be in line with the results of Broadbent and Gregory (1963), Loeb and Binford (1964), and Wiener, Poock, and Steele (1964, as reanalyzed by Taylor, 1965), it would conflict with those of Mackworth and Taylor (1963) and Mackworth (1965), where a decline in d' with time on task was consistently found. Although the experiments in which no change in signal detectability has been observed varied both in the sensory mode employed (visual or auditory) and in the nature of the task itself (discrimination between "signal" and "nonsignal" events, or detection of a change in a continuous input), the Mackworth studies are the only ones in which the particular combination of visual mode and continuous input has been investigated; this fact may be of significance for future research in this area.

REFERENCES

Broadbent, D. E., & Gregory, M. Vigilance considered as a statistical decision. *British Journal of Psychology*, 1963, **54**, 309–323.

Colquhoun, W. P., & Baddeley, A. D. Role of pretest expectancy in vigilance decrement. *Journal of Experimental Psychology*, 1964, **68**, 156–160.

Evans, G. W. Risk-taking set and target detection performance. *Journal of Applied Psychology*, 1965, **49**, 243–244.

Jerison, H. J., & Pickett, R. M. Vigilance: A review and re-evaluation. *Human Factors*, 1963, **5**, 211–238.

Loeb, M., & Binford, J. R. Vigilance for auditory intensity changes as a function of preliminary feedback and confidence level. *Human Factors*, 1964, **6**, 445–458.

Mackworth, J. F. Decision interval and signal detectability in a vigilance task. *Canadian Journal of Psychology*, 1965, **19**, 111–117.

Mackworth, J. F., & Taylor, M. M. The d' measure of signal detectability in vigilance-like situations. *Canadian Journal of Psychology*, 1963, **17**, 302–325.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. Decision processes in perception. *Psychological Review*, 1961, **68**, 301–340.

Taylor, M. M. Detectability measures in vigilance: Comment on a paper by Wiener, Poock and Steele. *Perceptual & Motor Skills*, 1965, **20**, 1217–1221.

Wiener, E. L., Poock, G. K., & Steele, M. Effect of time-sharing on monitoring performance; simple mental arithmetic as a loading task. *Perceptual & Motor Skills*, 1964, **19**, 425–440.

(Received May 26, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 2, 191–200

EMPIRICAL AND THEORETICAL LIMITATIONS OF THE TWO-FACTOR HYPOTHESIS OF JOB SATISFACTION

JOHN R. HINRICHS
Data Processing Division, IBM Corporation, White Plains, New York
AND LOUIS A. MISCHKIND
Systems Manufacturing Division, IBM Corporation, San Jose, California

The salient reasons for current job satisfaction were compared for high- and low-satisfaction respondents (613 technicians) in the framework of Herzberg's 2-factor hypothesis. An alternate notion stated that "motivators" are the prime influencers of satisfaction while "hygiene" factors act to limit complete satisfaction for highs and complete dissatisfaction for lows. The data, while strictly supporting neither formulation, were more adequately coordinated with the alternate conception. The mounting body of contradictory results and the inability of the 2-factor theory to handle deviant cases calls for a fresh look at the satisfiers/dissatisfiers concept. A new construct is offered in terms of Rotter's Social Learning Theory.

Since publication of *The Motivation to Work* (Herzberg, Mausner, & Snyderman, 1959), a flurry of articles and reviews have appeared which purport to throw light on the interpretation of job satisfaction advanced in that much-discussed book. Briefly stated, the notion proposes that two distinct sets of factors account for overall job satisfaction. The first set, called motivator factors or satisfiers, includes variables which are generally work related, for example, work interest and challenge, achievement, advancement, recognition, and so forth. The second set, termed hygiene factors or dissatisfiers, relate to variables which are work supporting or contextual to the work, including such things as technical supervision, corporate policies, interpersonal relationships, physical surroundings, etc. The core of Herzberg's concept maintains that motivator factors account for variance in overall job satisfaction above a neutral hedonic level with little or no bearing on satisfaction below that level. Hygiene factors, on the other hand, account for variance in overall job satisfaction below a neutral hedonic level with little or no bearing on overall satisfaction above that point. In short, the claim is that motivator variables operate primarily on the positive side of the overall job-satisfaction scale while hygiene variables operate on the negative side (Herzberg et al., 1959, pp. 111–112).

Despite several studies specifically designed to test the concept, not to mention numerous studies of general job attitudes which have been interpreted within the framework of the Herzberg formulation, at the present time it is not at all clear the extent to which the two-component concept of job satisfaction can or should be generalized beyond the specific population and methodology employed in the initial research.

Several studies have essentially replicated the Herzberg design and have yielded comparable results. Myers (1964) reported data from interviews with 282 employees at Texas Instruments, Inc., representing a range of job categories from scientists to female assemblers. Although there were some significant differences between job categories, the overall trends were interpreted as supporting *The Motivation to Work* explanation. Schwartz, Jenusaitis, and Stark (1963) also reported supportive findings from their questionnaire study among supervisory personnel in the utility industry. Herzberg (1965) presents comparable findings from data collected among supervisors in Finland.

However, as outlined by Ewen (1964) as well as by Vroom (1964), there are some basic questions surrounding the research

strategy and methodology utilized in these studies which should caution against too-quick acceptance of the interpretation given for the findings and their extension to the world of work in general. These cautions have been reinforced by data from a number of studies explicitly discussed within the framework of the Herzberg explanation which at least partially, if not totally, fail to support the predictions which one would make based upon the basic two-component hypothesis of job satisfaction (Ewen, 1964; Friedlander, 1963, 1964; Gordon, 1965; Halpern, 1965; Malinovsky & Barry, 1965; Wernimont, 1966).

Furthermore, in addition to the growing number of studies which raise important questions about the concept, it is believed that Herzberg's data do not adequately test his own notion because the research was not based solely on current satisfaction with a presently existing job situation. Herzberg (as did the other supportive studies cited above) asked his subjects (*Ss*) to think of a time when they were particularly happy or unhappy with their job, *whether it be their present job or any other job they may have had.* As a result, there is no control over the sampling frame for the data and no clear-cut basis for drawing inferences about the relative contribution of various job factors to overall job satisfaction.

A theory is only as powerful as its ability to accommodate deviant cases, and there is nothing in the Herzberg notion which provides a basis for explaining cases which do not conform to the simple two-factor dichotomy. The concept of expectancy is one component which is clearly missing from the formulation and which, as Vroom (1964, pp. 17–18) points out, must be included as an integral part of any scheme for predicting levels of motivation or of job satisfaction.

A recent study by Turner and Lawrence (1965), while not interpreted within the *Motivation to Work* framework, presents compelling data showing the importance of moderating variables of work-group culture in the dynamics of job satisfaction—for workers of one cultural background, jobs characterized as varied, complex, and demanding were associated with high job satisfaction; for workers of a different cultural background, such jobs tended to be associated with low job satisfaction in direct contrast with what one would expect from the simple motivator/hygiene concept.

It is contended, moreover, that still other factors must be included in any theory which hopes to explain the basic dynamics of job satisfaction. Certainly experience or exposure to various job attributes is one. The importance which an individual ascribes to different job attributes is another factor; an importance variable can reflect other critical factors such as personality variables and cultural or social system variables which are certainly to be accounted for in any predictive theoretical statement.

Most of these considerations are treated only tangentially, if at all, in the Herzberg analysis; at the end of this paper an explicit hypothesis of job satisfaction shall be developed based upon the Social Learning Theory (Rotter, 1954) which attempts to incorporate these missing components. But first, the study reported below may provide additional insights toward the development of a more comprehensive theory of job satisfaction.

The Present Study

The study reported in this paper investigated the tenability of the Herzberg et al. (1959) hypothesis concerning the satisfier/dissatisfier or motivator/hygiene effect on overall job satisfaction using data solely assessing *present* satisfaction with a *current* job situation. In addition, the study looked at respondent's perceptions of factors which tended to influence their current overall satisfaction positively, as well as their perceptions of factors which tended to influence satisfaction negatively.

The hypotheses for this study were directly at variance with what one would predict from the Herzberg et al. theoretical statement. Rather than operating unidirectionally, it was hypothesized that the types of job-content factors which Herzberg labeled "motivators" would operate bidirectionally by being both the primary cause of positive satisfaction in high-satisfaction respondents, as well as the

primary cause of negative satisfaction for low-satisfaction *Ss*. It was hypothesized that "hygiene" variables, on the other hand, would thus be responsible for the lack of total satisfaction for high-satisfaction *Ss* and for the lack of total dissatisfaction for the low-satisfaction *Ss*.

In other words, motivators were predicted to be prime movers of satisfaction. Because of the high degree of ego involvement associated with such job attributes as work challenge, achievement, recognition, advancement, and the other motivators, it was felt that these should logically represent the most powerful variables psychologically in contributing to job satisfaction, and conversely their absence should be most potent in contributing to job dissatisfaction. On the other hand, hygiene factors should logically operate as relatively weak influences; that is, the reasons satisfied people are not completely satisfied and the reasons dissatisfied people are not completely dissatisfied—"nit picks," if you wish.

Table 1 summarizes the prediction which would flow from the Herzberg hypothesis, and Table 2 highlights the relationships hypothesized for this study. Each table shows, for individuals who report they are currently satisfied and for those who report they are dissatisfied, the predicted most-salient factors which would be cited as having the most influence in raising current satisfaction and those which would be cited as having the most influence in depressing satisfaction.

Method

Respondents

The *Ss* for the present study were 613 technicians involved in service work all of whom were

TABLE 1

Predictions from Herzberg's Formulation of the Most Salient Factors Operating as Positive or as Negative Influences on Overall Job Satisfaction as a Function of Present Level of Satisfaction

Direction of influence	Level of present job satisfaction	
	Low	High
Positive factors	Primarily hygiene	Primarily motivators
Negative factors	Primarily hygiene	Primarily motivators

TABLE 2

Prediction for Present Study of the Most Salient Factors Operating as Positive or as Negative Influences on Overall Job Satisfaction as a Function of Present Level of Satisfaction

Direction of influence	Level of present job satisfaction	
	Low	High
Positive factors	Primarily hygiene	Primarily motivators
Negative factors	Primarily motivators	Primarily hygiene

presently employed by the same large national company. In the spring of 1965 a general attitude survey of the entire population of technicians was carried out. The survey resulted in a 75% response rate for a total of 10,000 completed questionnaires. The average age of the participating respondents was 32.7, with 92% of the sample under 45 years of age. There were 84% who had some education beyond a high school diploma. Nonmanagerial positions were held by 93%, and 85% had been with the company 3 years or more. All of the respondents were males. From the returned questionnaires, 1 out of every 16 were systematically selected to form the sample for the present study. The characteristics of the selected sample closely mirrored those of the population from which it was drawn.

Measuring Overall Job Satisfaction

At the beginning of the attitude-survey questionnaire an overall-job-satisfaction question was asked, utilizing an 11-point response scale:

> Now, considering everything, how would you rate your present satisfaction as an [XYZ Company] employee? (Circle one number on the scale below which best expresses how you feel.)
>
> 0 1 2 3 4 5 6 7 8 9 10
>
> Completely Dissatisfied — Completely Satisfied

Since this was the first question in the questionnaire it was essentially an unstructured assessment of the respondent's overall level of current satisfaction.

Measuring Job Factors Responsible for Overall Satisfaction

Immediately following the overall-satisfaction question were the following open-ended items:

> A. List one or two specific things that most influence your feelings in a *positive way* about your employment with [XYZ Company].
>
> 1. ______
> 2. ______

B. Now, list one or two specific things that most influence your feelings in a negative way about your employment with [XYZ Company].

1. ______
2. ______

High-Low Overall Job Satisfaction

Responses for the 613 men in the sample had a mean of 7.04 on the overall-job-satisfaction question and a standard deviation of 1.82. Respondents scoring 6 and below on the scale ($N=158$) were classified as forming a "low present satisfaction" group, and those scoring 7 or above ($N=455$) a "high satisfaction" group. The selection of the cutting point was dictated in part by the character of the open-ended responses; 5% of the respondents checking "7" indicated that they had "no negative feelings" about the job, whereas not a single respondent checking "6" failed to mention at least one negative reason. Since the intent was to separate high and low overall hedonic tone about the employment context, there was no clear-cut rationale for splitting respondents at the median on the satisfaction scale; rather, an attempt was made to find the point at which attitudes appeared to change from a positive to a negative tone. The authors were also constrained, however, to keep the cutting point as close as possible to the apparent midpoint of the scale. (The results discussed below are consistent across the full range of the satisfaction scale, becoming more pronounced the lower the cutting point.)

Classification of Satisfiers and Dissatisfiers

The two authors independently classified the two sets of open-ended responses. In order to reduce a hypothesis-confirming bias, neither had any knowledge of the respondents' actual satisfaction ratings. The classification categories are listed below along with comments where there was deviation from Herzberg's procedure. Where no comments are given, the descriptive scheme presented in *The Motivation to Work* (pp. 44–49) was followed.

Satisfiers:

(*a*) Recognition—This category includes general replies in which the respondent indicated he felt or did not feel respected for his worth as an individual, in addition to comments dealing with specific instances of recognition or its lack.
(*b*) Achievement
(*c*) Growth of Skills and Abilities
(*d*) Advancement
(*e*) Work Itself
(*f*) Autonomy

Dissatisfiers:

(*a*) Interpersonal Relationships with Peers and/or Subordinates—While 7% of this sample were managers, they were not identified as such when coding the open-ended questions. It was therefore difficult to tell whether the interpersonal relationship referred to a peer or to a subordinate. Therefore, these categories have been combined.
(*b*) Interpersonal Relationship—Supervisor
(*c*) Supervision—Technical
(*d*) Company Policy and Administration
(*e*) Working Conditions
(*f*) Job Security
(*g*) Benefits—This is not a Herzberg category and would ordinarily be classified with Corporate Policy. In this analysis, possibly because of the culture of the specific company investigated, it loomed as a highly distinct response.
(*h*) Status—Status responses per se were not obtained. Herzberg's scheme calls for responses referring to physical signs of status; status responses registered in this study dealt with the status derived from belonging to an organization like "the XYZ Company." These were eventually collapsed into Corporate Policy.
(*i*) Personal Life

Others:

(*a*) Wages—Since wages may be either satisfiers or dissatisfiers depending upon their implications, they were classified separately where the implication was not explicit.
(*b*) No Response—*No* negative or positive feelings were coded. If a respondent said, "I really do not have any negative (positive) feelings but . . .," it was coded as "no negative (positive) feelings." If a questionnaire had no responses in both the positive *and* negative categories it was deemed invalid. (1.9% of the questionnaires were considered invalid, reducing the final N to 613 respondents from an initial sample of 625.)

Reliability of the Coding

Considering all 16 coding categories, the rate of agreement between the two independent raters was 80.7%, and π, the index of intercoder agreement, was 78.7 (Robinson, 1957). The index reflects the extent to which coding reliability exceeds chance.

Considering only the three categories of "dissatisfiers," "satisfiers," and "wages," the rate of agreement was 91.6% with $\pi=84.3$.

For each and every disagreement, unanimous agreement was eventually reached between the two coders, and the finally agreed upon code was used in subsequent data analyses.

Weighting of Open-Ended Responses

Each respondent was instructed to list "one or two specific things" as negative and as positive reasons for job satisfaction. This was done to allow more than one job factor to emerge if it were sufficiently important. In some cases, two distinct factors were mentioned; in others, two aspects of the same factor were mentioned; and finally, some respondents mentioned only a single aspect of one job factor. In coding responses, each positive and each negative set of responses for each S was given

TABLE 3

LEVEL OF PRESENT OVERALL JOB SATISFACTION

Type of motivation factor	Low (N = 158)						High (N = 455)						Total group (N = 613)					
	Positive reasons		Negative reasons		Total		Positive reasons		Negative reasons		Total		Positive reasons		Negative reasons		Total	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Motivators	95	30.7	91	29.9	189	30.3	388	42.7	135	16.4	523	30.2	483	39.7	229	20.2	712	30.3
Hygiene	200	64.5	168	53.5	368	59.0	471	51.9	614	74.7	1085	62.7	671	55.1	782	68.8	1453	61.7
Money	15	4.8	52	16.6	67	10.7	49	5.4	73	8.9	122	7.1	64	5.2	125	11.0	189	8.0
Total	310	100	314	100	624	100	908	100	822	100	1730	100	1218	100	1136	100	2354	100
None Mentioned	6		2		8		2		88		90		8		90		98	

TABLE 4

ANALYSIS OF CHI-SQUARE OF DATA IN TABLE 3

Source	df	χ^2
Level of Job Satisfaction × Positive and Negative Reasons	1	1.5*
Level of Job Satisfaction × Type of Motivation Factor	2	8.4**
Type of Motivation Factor × Positive and Negative Reasons	2	115.9***
Type of Motivation Factor × Positive and Negative Reasons × Level of Job Satisfaction	2	56.7***
Total	7	182.5***

* $p > .10$.
** $p < .01$.
*** $p < .005$.

a weight of 2. Thus, for example, if two job factors were mentioned as positive reasons each received a weight of 1; if the same job factor was mentioned twice as a positive reason, it received a weight of 2; and if only one aspect of one job factor was mentioned it too received a weight of 2. This procedure in effect weighted responses for saliency and equalized the weight given to respondents with differing levels of verbalization. It also allowed for the inclusion of all factors mentioned. On the average, each *S* mentioned three and one-half separate responses; 59% of the *Ss* made two distinct responses in the positive and two in the negative categories. In those cases where more than two responses were given, only the first two were coded.

RESULTS

The primary data are presented in Table 3 and show the number of weighted mentions of the motivation factors and the proportions with which each were mentioned as positive or as negative reasons for the present level of overall satisfaction by the high- and low-satisfaction groups. A three-way analysis of chi-square was performed (Winer, 1962), and the results are presented in Table 4.

1. Level of Job Satisfaction × Positive and Negative Reasons. This result indicates that the proportion of positive versus negative reasons is not significantly related to level of job satisfaction. This is to be expected since the instructions and the weighting scheme forced two responses each as positive and negative reasons for each respondent. It is true, however, that a "no mention" was considered a valid response.

2. Level of Job Satisfaction × Type of Motivation Factor. This statistically signifi-

cant chi-square is accounted for by slight differences between the two groups in the frequency with which money and hygiene variables are mentioned.

3. Type of Motivation Factor × Positive and Negative Reasons. This significant result accounts for the greatest proportion of the total chi-square. It indicates that:

(*a*) A greater proportion of hygiene factors are mentioned as negative as opposed to positive reasons ($p < .0001$).

(*b*) A greater proportion of motivator factors are mentioned as positive as opposed to negative reasons ($p < .0001$).

(*c*) A greater proportion of money factors are mentioned as negative as opposed to positive reasons ($p < .0001$).

(It should also be noted that Hygiene factors account for significantly greater than 50% of the mentions as either positive or as negative reasons.) (p pos $< .005$; p neg $< .0001$)

4. Motivation Factor × Positive and Negative Reasons × Level of Job Satisfaction. This result is highly significant and bears most directly on the hypothesis of this study. To interpret this interaction it is recommended (Winer, 1962) that the table be split on the classification categories of one of the three variables, in this case the positive and negative reasons. It can be observed that the high-satisfaction group more often mentions motivators and less often mentions hygiene factors as positive reasons ($p < .005$) than does the low-satisfaction group. Money is mentioned as a positive reason equally frequently by both groups ($p < 1.00$). As negative reasons the low-satisfaction group more often mentions motivators ($p < .0001$) and less often hygiene factors ($p < .0001$) as well as more often mentioning money factors ($p < .005$) in comparison with the high-satisfaction group.

Figure 1 summarizes the results. Essentially the left-hand side of the figure displays the response pattern for the low-satisfaction respondents while the right-hand side shows the pattern for the high-satisfaction *S*s. The direction of the arrow indicates, for each satisfaction group, whether the predominant direction of the factor was statistically significant as a positive reason (arrow points right) or as a negative reason (arrow points left). The length of the arrow indicates the level of the predominant proportion. The width of the arrow reflects the relative frequency of mention of the factor for each satisfaction group. The differences in frequency of mention between the high- and low-satisfaction groups is also reflected in the width of the arrow. (Since the positive versus negative mention of motivators for the low-satisfaction group is not significantly different, left- and right-pointing arrows are drawn.)

Figure 1 indicates that motivators are predominantly influencing satisfaction positively

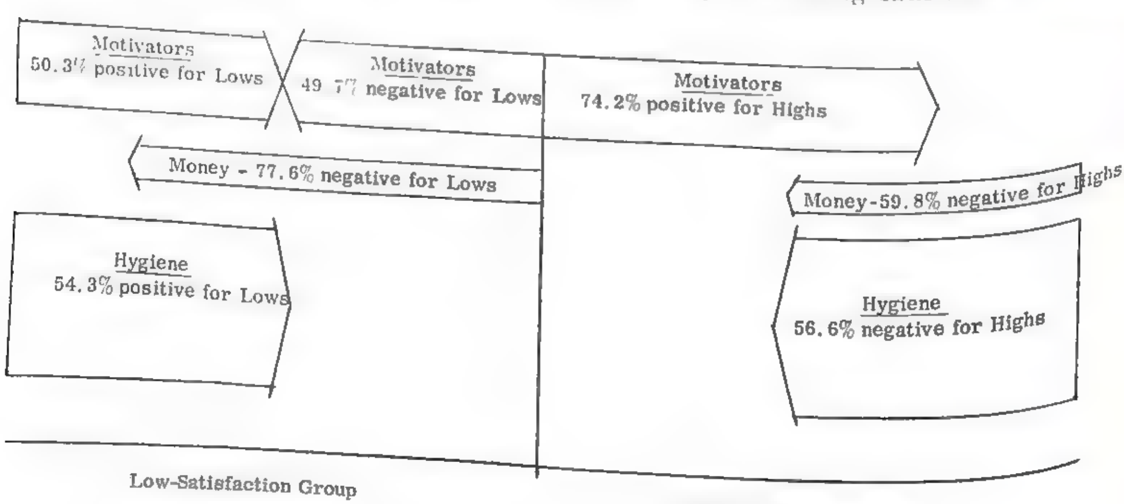


FIG. 1. Frequency and direction of mention for type of motivation factor by high- and low-satisfaction groups.

for the high-satisfaction group, while, for the low-satisfaction group they have equal positive and negative influence. Hygiene factors act predominantly negatively for the high-satisfaction group (as reasons for incomplete satisfaction) and predominantly positively for the low-satisfaction group (as reasons for incomplete dissatisfaction).

Discussion

The observed data confirm neither the Herzberg et al. (1959) hypothesis nor the full hypothesis set for this study. One result which significantly affects the data is the extent to which hygiene factors are mentioned more frequently than motivator factors by both high- and low-satisfaction *Ss*, both as negative and as positive reasons. Four possible explanations are offered for this:

1. A Method Factor Unique to This Study. Perhaps respondents attended to the phrase "XYZ Company" in the stem of the open-ended question. If they interpreted this as calling for company-policy-like responses, a spurious frequency of hygiene factors may have resulted.

2. A Method Factor Unique to Herzberg's Study. Perhaps asking an *S* to think back to some time when positive or negative feelings were intense elicits a spurious frequency of socially acceptable "important" factors, that is, motivators.

3. The Job Category of the Respondents. Perhaps technicians deem hygiene items relatively more salient than accountants or engineers.

4. The Company of the Respondents. The company in which the respondents were employed is known to have a highly comprehensive benefit program in addition to a well-developed system of employee-higher-management communication and concern for the maintenance of high morale. Perhaps the relative saliency of hygiene and motivator factors within this particular culture is influenced by the nature of company practices.

While neither hypothesis explicitly takes into account the possibility of markedly unequal frequencies of hygiene and motivator factors, they both do permit some prediction about the relative frequency of these two classes of variables.

Herzberg's scheme would predict that motivators would be relatively more frequent as positive reasons for high-satisfaction *Ss* than as positive reasons for low-satisfaction *Ss*. The present hypothesis would make the same prediction. The data support this prediction.

Herzberg's hypothesis would predict that hygiene variables would be relatively more frequent as negative reasons for low-satisfaction *Ss* than as negative reasons for high-satisfaction *Ss*. The present hypothesis predicts the reverse. The data support the present hypothesis. If hygiene factors are the culprits in determining negative satisfaction below a neutral point as one would predict from the Herzberg formulation, then dissatisfied, not satisfied, employees should mention a greater proportion of them as negative reasons.

The present formulation implies four other explicit predictions about the relative frequency of positive and negative reasons:

1. Hygiene variables should be mentioned relatively more often as negative than as positive reasons by high-satisfaction *Ss*.

2. Hygiene variables should be mentioned relatively more often as positive than as negative reasons by low-satisfaction *Ss*.

3. Motivators should be mentioned relatively more often as positive than as negative reasons by high-satisfaction *Ss*.

4. Motivator variables should be mentioned relatively more often as negative than as positive reasons by low-satisfaction *Ss*.

Predictions 1 through 3 were confirmed but Prediction 4 was not; there was no significant difference observed. (Herzberg's hypothesis only touches on these issues and does not make any explicit predictions about these relationships.)

Myer's (1964) interpretation extends the Herzberg hypothesis by maintaining that hygiene variables "have little motivational value . . . [p. 85]" but only became important in determining below-neutral variance in satisfaction when motivators are absent from the job environment. It is as if the *absence* of

TABLE 5

SATISFACTION-SCALE RESPONSE VERSUS FREQUENCY OF MOTIVATORS

Satisfaction-scale response	Completely dissatisfied 0–3	4–5	6	7	8	Completely satisfied 9–10	Total
Number of respondents	28	78	52	147	210	98	613
Frequency of mention of motivators	29%	29%	32%	31%	27%	28%	29%
Proportion of motivators in negative direction	59%	51%	43%	37%	22%	16%	32%

motivators were not attended to as a source of dissatisfaction by low-satisfaction *Ss*. It is not clear whether the absence is suppressed, repressed, or loses saliency. (There is implied in Myer's discussion a subtle distinction between the *absence* of motivators as opposed to their dissatisfaction, that is, the distinction between achieving versus standing still and between achieving versus failing. This issue is an extremely important one inadequately dealt with by all hypotheses including the present one.)

If motivators are increasingly repressed, suppressed, or less salient as satisfaction decreases, their overall frequency of mention should decline, and particularly their frequency of mention as negative reasons for satisfaction. However, it is evident from Table 5 that the frequency of motivators does not decline as satisfaction decreases. In addition, the proportion of motivators mentioned in the negative direction becomes greater as satisfaction decreases, a finding directly at variance with the loss-of-saliency hypothesis.

The primary purpose of the present study was to test the generality of Herzberg's hypothesis when overall satisfaction with a presently existing job situation is assessed. On the basis of these and other contradictory results, there is doubt about the wisdom of the simple two-factor approach for describing the determinants of overall job satisfaction. What has become particularly apparent is the significant frequency of deviant cases with patterns of response which do not conform to any hypothesis yet proposed.

It is evident that the issues which Herzberg raises are of extreme importance. Unfortunately, however, these basic issues have been spelled out rather vaguely and have been confused by many investigators. While Herzberg's basic findings were that content/context factors were correlated with satisfied/dissatisfied attitudinal responses, the extension of his data to the two-factor notion assumes a direct equivalence between the two.

On logical grounds there are four possible patterns of relationship between the saliency of a job factor and satisfied/dissatisfied attitudinal responses:

1. A factor may be salient when an individual states that he is satisfied, but not when dissatisfied.
2. It may be salient when dissatisfied, but not when satisfied.
3. It may be salient both when satisfied and when dissatisfied.
4. It may not emerge as salient either when the individual is satisfied or when he is dissatisfied.

Herzberg interprets his data as indicating that motivators fit the first pattern and hygiene variables the second pattern. The construct he employs to explain this is based primarily on a moral definition of the nature of man and an analogy (Human beings are essentially characterized as self-actualizing organisms and anything which is not a self-actualizer acts as a hygienic factor). This construct seems a trifle ad hoc and is not coordinated to more basic psychological variables.

A more powerful theory would provide for the prediction of all possible satisfier/dissatisfier patterns for any given job factor in any cultural setting. It is felt that a variation of the Social Learning Theory (Rotter, 1954) meets these requirements.

The basic function would state:

$$HV = f\left[RV\left(\frac{Occ - Exp}{N}\right)\right]$$

where: HV = Hedonic Value: an index of the saliency of a specific factor in a positive or negative direction. Thus, in effect this would be an index of the extent to which the factor serves as a satisfier or dissatisfier. RV = Reinforcement Value or Valence: basically a measure of the relative importance of the factor with respect to other factors in the individual's system of work goals, where: $-1.0 \leq RV \leq +1.0$. Occ = The extent of the occurrence of reinforcement for a factor: this may be stated as a decimal which in instances of a single reinforcement can take on values of 0 or 1, and where $-1.0 \leq Occ \leq +1.0$. Occ is a reflection of the individual's prior experience with regard to the particular job factor. Exp = The subjective probability of reinforcement for the factor: $-1.0 \leq Exp \leq 1.0$. This is essentially an expectancy variable. N = Some function of the frequency or number of previous experiences with regard to the job factor under consideration. As the inequalities indicate, RV, Occ, and Exp can logically accommodate factors viewed in a bipolar fashion by taking on either positive or negative signs.

Consider as an example the motivator of advancement or promotion in the organization. In terms of this predictive function, if an individual views his probability of advancement to lie on a continuum from no advancement to advancement, his expectations or Exp would be positive. If, on the other hand, he sees his probability of advancement to lie on a continuum from no advancement to demotion, Exp would be negative.

Similarly, Occ would be positive if in fact he were promoted, that is, if the reinforcement fell on the continuum from no advancement to advancement. Occ would be negative if the reinforcement were from no advancement to demotion.

Finally, if promotion represented a valued goal for the individual, RV would be positive; if he viewed a promotion as undesirable or threatening for some reason, RV would be negative.

In this particular example of promotion or advancement, N might possibly be viewed as a function of length of time since last promotion, such that, for a short period since an individual's last promotion N would be large, and for a long period N would be small. An expression such as $N = \dfrac{1}{\text{years since last promotion}}$ might adequately describe the variable in this instance. This merely accommodates the fact that valued reward for the "hungry" man has more impact than reward for the satisfied individual.

By inserting various assumed values for these variables into the function, one can identify those sets of conditions under which a particular work-content or -context factor can be expected to serve as a satisfier, and when as a dissatisfier. This type of model permits prediction to the satisfied/dissatisfied dimension from pertinent basic psychological variables associated with the content/context factors. Table 6 summarizes the predictions which would flow from most of the pertinent permutations of the variables in the model. Assuming values of +1, 0, or −1 for RV, Occ, and Exp, and assuming the N term to be a constant equal to 1, HV could range in value between −2 and +2—the "motivator" of advancement could conceivably range from being a strong dissatisfier ($HV = -2$) through a position of indifference ($HV = 0$) to a

TABLE 6

SIGNIFICANT PERMUTATIONS OF RV, Occ, AND Exp

Possible values			Resultant HV	Explanation
RV	Occ	Exp		
+1	+1	−1	+2	Unexpected good fortune
−1	−1	+1	+2	Unexpected good fortune
+1	+1	0	+1	Gratuitous reward
−1	−1	0	+1	Gratuitous reward
+1	0	−1	+1	Unfounded apprehensions
−1	0	+1	+1	Unfounded apprehensions
±1	+1	+1	0	A completely predictable outcome
±1	0	0	0	No experience
0	±1	±1	0	Indifference
±1	−1	−1	0	A completely predictable outcome
+1	0	+1	−1	Unfulfilled expectations
−1	0	−1	−1	Unfulfilled expectations
+1	−1	0	−1	Unexpected loss
−1	+1	0	−1	Unexpected loss
+1	−1	+1	−2	Complete reversal in expected valued outcome
−1	+1	−1	−2	Complete reversal in expected valued outcome

strong satisfier ($HV = +2$), depending upon the importance an individual attaches to promotion, whether or not promotion occurs, and his prior expectations of advancement.

In this example, the value of

$$N = \frac{1}{\text{years since last promotion}}$$

would inflate the value of HV for individuals with a relatively long period of time since last reinforcement and reduce it for the recently promoted man. This experience variable is intuitively an important component of job satisfaction which is omitted from most models.

Beyond the example dealing with advancement, a similar rationale could be applied to any of the other content or context factors discussed by Herzberg, as well as to additional components of the work environment.

Although the validity of this specific model has not yet been empirically demonstrated, it is amenable to test, and data are currently being collected within this framework. Whether or not this exact function explains industrial job-satisfaction data, it is hoped that this discussion has highlighted some of the pitfalls in assuming a simple equivalency between content/context factors and satisfied/dissatisfied response patterns and that it presents some of the more basic variables which have to be considered. In short, it is believed that the relationship between reward value, reward occurrence, reward expectancy, and frequency of similar experiences will better predict the positive or negative influence of individual job factors on overall job satisfaction.

REFERENCES

EWEN, R. B. Some determinants of job satisfaction: A study of the generality of Herzberg's theory. *Journal of Applied Psychology*, 1964, **48**, 161–163.

FRIEDLANDER, F. Underlying sources of job satisfaction. *Journal of Applied Psychology*, 1963, **47**, 246–250.

FRIEDLANDER, F. Job characteristics as satisfiers and dissatisfiers. *Journal of Applied Psychology*, 1964, **48**, 388–392.

GORDON, G. G. The relation of "satisfiers" and "dissatisfiers" to productivity, turnover, and morale. Paper read at American Psychological Association, Chicago, September 1965.

HALPERN, G. Relative contributions of motivator and hygiene factors to overall job satisfaction. *Journal of Applied Psychology*, 1966, **50**, 198–200.

HERZBERG, F. The motivation to work among Finnish supervisors. *Personnel Psychology*, 1965, **18**, 393–402.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. B. *The motivation to work*. New York: Wiley, 1959.

MALINOVSKY, M. R., & BARRY, J. R. Determinants of work attitudes. *Journal of Applied Psychology*, 1965, **49**, 446–451

MYERS, M. S. Who are your motivated workers. *Harvard Business Review*, 1964, **42**(1), 73–88.

ROBINSON, W. S. The statistical measure of agreement. *American Sociological Review*, 1957, **22**, 17–25.

ROTTER, J. B. *Social learning and clinical psychology*. New York: Prentice Hall, 1954.

SCHWARTZ, M. M., JENUSAITIS, E., & STARK, H. Motivational factors among supervisors in the utility industry. *Personnel Psychology*, 1963, **16**, 45–53.

TURNER, A. N., & LAWRENCE, P. R. *Industrial jobs and the worker*. Boston: Harvard University Graduate School of Business Administration, 1965.

VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.

WERNIMONT, P. F. Intrinsic and extrinsic factors in job satisfaction, *Journal of Applied Psychology*, 1966, **50**, 41–50.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received May 31, 1966)

journal of applied psychology

VOL. 51, No. 3 JUNE 1967

THE *K* FACTOR AS A (VALIDITY) SUPPRESSOR VARIABLE IN PREDICTING SUCCESS IN SELLING

FLOYD L. RUCH
University of Southern California

AND

WILLIAM W. RUCH
Psychological Services, Inc., Los Angeles

The MMPI was given to 182 sales representatives from 9 companies, ranging from beverage sales to business forms. Within each sample, Ss were cast into upper- and lower-criterion groups, based upon sales managers' ratings. The 5 uncorrected MMPI scales which are normally subject to the K correction differentiated significantly between the criterion groups in the expected direction. The K suppressor variable, when applied as instructed in the publisher's manual, actually decreased validity. This effect was clearly statistically significant for 3 of the scales, approached significance for 1 scale, and was not significant for the final scale.

The senior author of this paper has been given credit (Meehl & Hathaway, 1946) for being the first psychologist to construct a "lie detector" scale to measure faking on a pencil-and-paper self-inventory test of personality (Ruch, 1942). Although this original article has been overlooked by most psychologists because of the obscure place of its publication, it did serve as a prototype for the development of the *K* scale for the Minnesota Multiphasic Personality Inventory (MMPI). The purpose of the present study was to investigate the validity of the *K* scale as a suppressor variable in the employment situation.

Despite the criticisms of numerous journalists and a few psychologists, self-inventory tests and adjective checklists *without corrections for faking* have been found to correlate with success on the job—especially in selling (Dunnette, McCartney, Carlson, & Kirchner, 1962; Ghiselli & Barthol, 1953; Husband, 1936), and more especially when specialized keys are developed through item analysis for particular jobs (Guion & Gottier, 1965). The findings to be reported in this paper are in agreement with those just cited.

There are two possible explanations for this moderate but economically valuable predictive power of the self-inventory: (*a*) these instruments work *despite* the fact that they can be faked; or (*b*) they work *because* they can be faked. The present authors are strongly convinced that the weight of the evidence is for the latter explanation. This needs some amplification. It is the authors' hypothesis that the good salesman is more likely than the poor salesman to have a clear conception of what demands the selling job puts on him personality-wise, and hence is better able than the poor salesman to put his best foot forward, regardless of what his "true" personality dynamics really are. The present authors have named this hypothetical construct "JID," *job-image discrepancy*. Their contention is that salesmen, especially, will attempt to give the answers they think the employer wants, and hence their responses become to a large degree a measure of their image of the job's demands. This is akin to what Gellermann (1963) has called "sensible deception."

However, the "sensible deception" concept as used by Gellerman to describe what happens in the real-life situation is not the same as an all-out attempt to fake when subjects (*Ss*) are instructed to do so as part of an experiment "that doesn't count" (Dunnette et al., 1962).

What, then, does *K* measure for job applicants? In reviewing the literature, Dahlstrom and Welsh (1960) conclude that:

> The picture [of high *K* normals] is one of poise, reserve, and comfort in social relations. These people usually have the social skills and experience to carry off successfully the moment-to-moment requirements of a social situation, or if they have qualms and insecurities, to cover them under a façade of imperturbability [p. 144].

This statement could easily be a description of the successful salesman!

The *K* scale is also a measure of the extent to which a testee gives socially acceptable responses. In the normal counseling situation, there will be a wide range of test-taking attitudes. Some *Ss* will be extremely honest and frank; some will be extremely defensive. By measuring this attitude with the *K* scale and correcting the clinical scales for this test-taking attitude the authors of the MMPI feel that the discriminatory power will be sharpened (Hathaway & McKinley, 1951). However, it is reasonable to assume that the variability in test-taking attitude among job applicants is much less than among counseling cases. The former are motivated to "look good" to the employer. It is hypothesized that the good salesman will succeed better in giving the "right" answers because his job image is less discrepant. If this hypothesis is true, the function of *K* correction would be to suppress *validity* rather than test-taking attitude.

In the present study, the specific hypothesis was that the *K* correction will reduce the validity of the five clinical scales to which it is applicable. A presupposition, also tested, was that the uncorrected scales are correlated with selling effectiveness.

METHOD

Subjects. The sample consisted of 182 presently employed sales representatives from nine different organizations. All were in industrial rather than consumer sales; that is, the customers were businessmen, not members of the general public.

Procedure. As part of a larger battery, the MMPI was administered to *Ss* while they were employed with their respective companies. Although the primary purpose of the testing was to establish standards against which to evaluate future applicants, *Ss* were told that test results would be reported to management, as indeed they were. It was felt that this was sufficient motivation to cause *Ss* to answer the questions as they would, had they been applicants. Prior to taking the battery, *Ss* read a sheet entitled, "Why Take Tests?" which briefly described the purpose of testing. Part of the description of the temperament tests was as follows:

> In taking tests of this sort, it is particularly important to describe yourself as accurately as you can. People have a natural tendency to "put their best foot forward." This is especially true of salesmen, who quite properly sell themselves as well as their products or services. If this tendency is too extreme, the results will be difficult to interpret. For this reason the tests contain a number of cross checks. These and other "lie detector" features determine the extent to which a person has tried to "beat the tests."

Before test results were reported to management, *Ss'* superiors rated them into upper and lower halves, based upon total effectiveness in their present jobs.

ANALYSIS AND RESULTS

The analysis was based on the five MMPI scales which are usually adjusted by adding a proportion of the *K* score. These scales are: Hypochondriasis, Psychopathic Deviate, Psychasthenia, Schizophrenia, and Hypomania. The keying is such that high scores are indicative of psychopathy.

The first question to be answered was, "Do these scales, without the *K* correction, differentiate between good and poor salesmen?" As parametric tests do not apply to these markedly skewed distributions, the Mann-Whitney *U* test was employed. The *Ss* were pooled into two groups—better and poorer salesmen. A one-tailed test was used, the null hypothesis being rejected if the better salesmen had significantly lower (more normal)

TABLE 1

RESULTS OF MANN-WHITNEY *U* TEST OF DIFFERENTIATION BETWEEN GOOD AND POOR SALESMEN FOR UNCORRECTED MMPI CLINICAL SCALES

Scale	z
Hypochondriasis	1.67*
Psychopathic Deviate	3.21***
Psychasthenia	2.48**
Schizophrenia	2.67**
Hypomania	1.89*

Note.—$N = 182$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

TABLE 2

COMPARISON OF RELATIVE PERFORMANCE ON CORRECTED AND UNCORRECTED SCALES FOR GOOD AND POOR SALESMEN

Scale	Criterion	Better (lower) on uncorrected	Better (lower) on corrected	χ^2	Direction
Hs	Good	56	30	4.7**	Uncorrected scale more valid
	Poor	45	49		
Pd	Good	60	27	3.2*	Uncorrected scale more valid
	Poor	51	42		
Pt	Good	51	34	4.5**	Uncorrected scale more valid
	Poor	40	53		
Sc	Good	56	32	9.8***	Uncorrected scale more valid
	Poor	36	56		
Ma	Good	38	50	0.6	Uncorrected scale more valid
	Poor	34	59		

Note.—Abbreviations: *Hs* = Hypochondriasis; *Pd* = Psychopathic Deviate; *Pt* = Psychasthenia; *Sc* = Schizophrenia; *Ma* = Hypomania.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

scores. As can be seen in Table 1, all five scales, *if uncorrected*, differentiated between better and poorer salesmen at accepted significance levels.

The second question to be answered was, "Does the *K* factor, when used to correct these five scales as recommended in the publisher's manual, increase or decrease their predictive power?" If the correction indeed increases validity, the better salesmen would have relatively lower scores and the poorer salesmen relatively higher scores after the correction was applied. To test this, the entire sample was ranked on each scale, both with and without the *K* correction. For each of the five scales, a four-fold table was constructed by splitting each of the criterion groups into those who did relatively better (that is, scored lower) on the corrected scale and those who did relatively better on the uncorrected scale. Table 2 shows these tables, chi-squares, significance levels (two-tailed tests), and the directions of significance. For all five scales, the *K* correction actually *decreased* validity. Using accepted levels, this was statistically significant for three scales, approached significance for one scale, and was not significant for the fifth.

The final question to be answered was, "What is the *degree* of correlation between the predictors and the criterion?" Due to the marked skewness of several of the distributions, all predictors were dichotomized and tetrachoric correlations computed. Since the criterion was established within subsamples (that is, within each company), the predictors were also dichotomized within subsamples, as nearly as possible, at their subsample medians. The *Ss* were then pooled into four-fold tables. Tables 3 gives the validity coefficients (tetrachoric correlation) for *K* and for the corrected and uncorrected clinical scales. It can be seen that most of the validities of the uncorrected scales are high enough to be useful as part of a battery. The *K* correction, however, reduces these validities to the point where they are essentially zero.

TABLE 3

TETRACHORIC VALIDITY COEFFICIENTS

Scale	Uncorrected	Corrected
K	.39*	—
Hypochondriasis	−.20	.02
Psychopathic Deviate	−.41	−.18
Psychasthenia	−.27	−.02
Schizophrenia	−.24	.10
Hypomania	−.10	−.07

Note.—$N = 182$.
* $p < .001$ (two-tailed test). Other significance levels are not reported, as they have been evaluated elsewhere.

REFERENCES

DAHLSTROM, W. G., & WELSH, G. S. *An MMPI handbook.* Minneapolis: University of Minnesota Press, 1960.

DUNNETTE, M. D., MCCARTNEY, J., CARLSON, H. C., & KIRCHNER, W. K. A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology,* 1962, **15**, 13–24.

GELLERMAN, S. W. Personnel testing: What the critics overlook. *Personnel,* 1963, **40**(3), 18–26.

GHISELLI, E. E., & BARTHOL, R. P. The validity of personality inventories in the selection of employees. *Journal of Applied Psychology,* 1953, **37**, 18–20.

GUION, R. M., & GOTTIER, R. F. Validity of personality measures in personnel selection. *Personnel Psychology,* 1965, **18**, 135–164.

HATHAWAY, S. R., & MCKINLEY, J. C. *Minnesota Multiphasic Personality Inventory manual.* New York: Psychological Corporation, 1951.

HUSBAND, R. W. Personality traits of salesmen. *Journal of Psychology,* 1936, **1**, 223–233.

MEEHL, P. E., & HATHAWAY, S. R. The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology,* 1946, **30**, 525–564.

RUCH, F. L. A technique for detecting attempts to fake performance on a self-inventory type of personality test. In Q. McNemar & M. A. Merrill (Eds.), *Studies in personality.* New York: McGraw-Hill, 1942. Pp. 229–234.

(Early publication received February 6, 1967)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 205-210

DEVELOPMENT OF A METHOD OF PREDICTING HIGH-ACCIDENT AND HIGH-VIOLATION DRIVERS [1]

BRUCE D. GREENSHIELDS [2]
University of Michigan

AND

FLETCHER N. PLATT [3]
Ford Motor Company

This research has been carried out to test the hypothesis that drivers with different driving histories (the high-accident, the low-accident, the high-violator, and the beginning driver) exhibit measurably different characteristics when tested on the same route in an instrumented car. The device used in collecting data, called a Drivometer, records in digital form: (a) driver control actions, (b) vehicle motions, and (c) traffic events. The measurements were taken over a route of approximately 17 mi. and revealed significant differences between the populations and reliability in individual prediction of about 2 in 3. Considering the chance factor involved in driving risks, this technique offers an important step in driver testing.

Methods of predicting system reliability, particularly in aerospace science, have advanced rapidly in the last few years. On the other hand, prediction of accident involvement, in regard to our 99 million plus licensed drivers, has experienced relatively little progress.

The reasons for this lack of progress are readily apparent. The analysis of a typical driving situation involves such a multitude of variables that their correlation in one instance does not necessarily hold true for the next. Although the highway system includes the vehicle, driver, and highway, the driver presents the greatest challenge in predicting accident involvement.

In order to gain a better understanding of the driver, it is logical to begin by examining objectively the behavior of the individual driver in the real traffic environment.

By first preselecting the drivers for this type of examination, admittedly a difficult task, and then following a systems approach to the complex problem, it has been possible to identify and measure the elements which relate to the driver and his environment. It has been found that certain measured driver controls, and related vehicle and traffic dynamics, are meaningful, and that group and individual classifications of drivers may be obtained by multivariate analysis of these measurements. This study is based on the hypothesis that drivers with different accident experience and driving records exhibit different driving profiles in ordinary driving.

An earlier study (Greenshields, 1963) suggested that it is possible to record and identify driving patterns and hence classify drivers. This study is essentially a repetition of the first, conducted with more reliable equipment and more refined techniques.

Test Procedures

Testing procedures were established on the assumption that drivers have certain basic characteristics related to their abilities and their physical and psychological traits. It was recognized that a driver's performance undoubtedly changes from time to time, but it was believed that drivers with proper instruction, designed to not only inform them but put them at ease, would perform essentially at their norm. The good student passes his examinations with consistently high grades. The superior athlete is still superior even during a "slump."

Methods of Recording Events

Events recorded were:

1. Vehicle in front of test car
2. Total traffic density around the vehicle:
   (a) Vehicle in front of test car.
   (b) Car passing on left

[1] This study was made possible by a grant from the Ford Motor Company Fund.

[2] Prior to June 1966 was Acting Director, Transportation Institute, and lecturer in Transportation Engineering, Department of Civil Engineering, University of Michigan.

[3] Manager, Traffic Safety and Highway Improvement Department, Ford Motor Company.

(c) Car passing on right.
(d)–(n) Pedestrians crossing, etc.
3. Total trip time.
4. Running time (car in motion).
5. Delay time—total time minus running time.
6. Amount of speed change.
7. Amount of direction change.
8. Acceleration reversals.
9. Brake applications.
10. "Gross" steering-wheel reversals.
11. "Micro" steering-wheel reversals.
12. "Micro" steering-wheel reversal rate.

All traffic and highway events, whether affecting the driving task or not, fell into two general classifications: fixed and variable. The fixed events included all stationary features such as curbs, signs, stop lights, trees, and buildings. Since all drivers traversed the same route the fixed events were not recorded. The variable events included all vehicles and pedestrians whether moving, standing, or parked.

In the enumeration of events, it was reasoned that a driver pays more attention to objects near him than to those farther away. A vehicle directly in front of the test vehicle may be noted, and also a second car might be noted as a discrete event, but more cars than that may be observed as a string of vehicles. One or several cars parked on the right or left side of the street are perceived simply as parked cars, or one event. An observer (*O*), using a hand-held keyboard, recorded the duration of events by depressing coded switch keys. For example, in response to the appearance of a car in front of the test car, *O* pushed the proper switch to the "on" position and left it there as long as the car was in front. A counter enumerated the seconds the car was in front. Thus, the number of keys depressed at any one time reflected the density of traffic events. The total time all of the keys were depressed was divided by the running time to give the average density of events for the trip.

The *O* was instructed to use his best judgment about how close a car in front should be in order to be considered an event worth recording. This distance increases directly with speed. The *O* was instructed, in case of doubt, to include the vehicle. Other events were similarly recorded.

All of the data on driver actions, vehicle motions, etc., were recorded on digital counters and the counters photographed at selected intervals.

Time variables were recorded by means of a constant-speed motor furnishing an electric impulse to counters once a second. The times recorded were total trip time, running time (time car is in motion), delay time (total time minus running time); the amount of speed change was recorded by electric impulses furnished by an auxiliary speedometer mounted under dash. The impulses were supplied at 2-mph intervals for either an increase or decrease in speed. Amount of direction change was obtained by electric impulses supplied from a gyro compass to a counter for each $2\frac{1}{2}°$ change in direction.

An accelerator reversal amounting to $\frac{1}{8}$ in. up and down movement from any position was recorded. The number of brake applications was recorded. The "macro" steering-wheel reversals were those amounting to about $8\frac{1}{2}°$ of turn of the steering wheel. This is about the amount of turning necessary to produce a noticeable change in direction of a car. The "micro" steering-wheel reversals were those amounting to $2\frac{1}{2}°$ of turn of the steering wheel.

In the collection of data, care was taken to follow the same procedure on all test runs. Instructions to the drivers were read in order that they would not vary. Only one research assistant accompanied each driver. Every effort was made to put the driver at ease. He was not taking a test in the ordinary sense for there would be no reward and no penalty. He was to drive over the speed limit if he habitually drove that way.

The runs were all made in good weather with dry pavement in a 1962 car (Ford Fairlane) with automatic shift, during the midmorning or midafternoon hours when the traffic flow was reasonably constant. The *O*, however, recorded the traffic density.

Drivers were tested in each class throughout the months taken to complete the field work. This was done in order to eliminate the effects of any minor changes in vehicle characteristics, route, and seasonal changes. However, a larger percentage of the beginning-driver runs were made during the first months of testing than were the other groups because of scheduling problems.

TEST ROUTE

The test route selected was 17.0 mi. long consisting of approximately 6.4 mi. of rural, 5.1 mi. of expressway, 2.0 mi. of commercial (downtown), 2.0 mi. of industrial, and 1.5 mi. of (high-density) residential. The average running time was approximately 28 min., about equally divided in time for each type of roadway.

SELECTION OF DRIVERS

To establish a standard of measurement, drivers with better than average records were selected from the lists supplied by an insurance agent. In addition to these, a group of driver-training instructors was screened; only those with the best driving records were used. These two groups were combined to furnish the control group (CG) composed of 17 driver-education teachers and 23 low-accident experienced drivers.

A sample of 40 drivers (high accident, HA) in jeopardy of losing their insurance because of too many accidents was selected at random from the files of an insurance agent. Individuals with more than three violations were eliminated to avoid confusing with the high-violation (HV) group.

A high-violation (HV) group was also drawn. Twenty drivers with over 12 Michigan points acquired in 2 yr. (sufficient for revocation of

TABLE 1

DESCRIPTION OF GROUPS—AVERAGES

Description	Control group (40 drivers)	High accident (40 drivers)	High violation (20 drivers)	Beginning (40 drivers)
Age	37.8	43.2	25.0	16.5
Percentage of males	80%	80.0%	90.0%	55.0%
Driving experience (in years)	21.8	22.9	10.4	$2\frac{1}{3}$ mo.
Miles driven last year	14,250	15,750	18,500	
No. accidents in 4 yr.	0.0625	3.0	0.85	
Violations (in points) over 4 yr.	0.5	1.7	13	
Percentage of zero violations	80	60	0	

license) were selected. Their accident records, however, were not above average.

Forty beginning drivers (BD) who had passed high school driver education and had their Michigan driver's licenses not more than 3 mo. were also selected.

An independent sample of drivers was selected at random from the four basic groups and not used in the primary analyses. The independent sample was used to test the validity of the discriminant-function coefficients and probability of predicting the correct classification of a single driver.

Information on age, sex, driving experience, etc., for these groups is shown in Table 1.

ANALYSIS OF DATA

The object of the analysis was to find some function of a set of driving variables that would discriminate between drivers of the different classes. The particular multivariate-analysis method employed is known as "discriminatory analysis [Bennett & Franklin, 1954]." (Computer Program Bimed. No. 005 UCLA.)

In analyzing the data, since the variables were not all entirely independent of each other, it was reasonable to suspect that a combination of part of them could be more effective in discriminating between the sets of individuals than a combination of all of them. For this reason different combinations were tested. The method followed consisted in using all of the variables for a first trial and then omitting one variable at a time until best results were obtained. In several cases fewer variables gave as good or better a separation of groups as did a larger number. For example, in discriminating between the high-accident (HA) groups and the control group (CG), four variables gave as good a discrimination between groups as did six variables.

To illustrate using Variables 2, 4, 8, 10, 11, and 12 to discriminate between the two groups, they were found to be different at a level of confidence of 97.5% (level of significance 2.5%). Using Variables 2, 4, 8, 10, and 11, the level of confidence was 99.0%.

The results of analyses runs to determine levels of significance of discrimination between groups and the number of variables used to give maximum levels of significance are shown in Table 2.

Comments:

(*a*) It is to be noted that the five variables that have high significance in the comparisons between any two groups are:

2. Traffic density. 4. Running time. 8. Accelerator reversals. 10. Gross steering reversals. 11. Fine steering reversals.

(*b*) For discrimination of the HA group or the HV group compared to the drivers with

TABLE 2

SIGNIFICANT LEVELS OF DIFFERENCES IN DRIVING BEHAVIOR

	F	p
CG versus HV Variables 4, 8, 10, 11	$F(4/55) = 3.5207$	2.5
CG versus HA Variables 4, 8, 10, 11	$F(4/75) = 8.4817$	0.1
HV versus BD Variables 4, 8, 10, 11	$F(4/55) = 6.2898$	0.1
HV versus HA Variables 2, 4, 8, 10, 11	$F(5/54) = 2.3090$	10.0
HA versus BD Variables 4, 8, 10, 11	$F(4/75) = 15.0182$	0.1

Note.—Abbreviations: CG = control group; HV = high violation; HA = high accident; BD = beginning drivers.

TABLE 3

PERCENTAGE OF DRIVERS CORRECTLY GROUPED BY RANKING

Group	N	Group	N	Percentage correctly placed
Control	40	Beginning	40	67.5%
Control	40	High violation	20	76.7%
Control	40	High accident	40	72.5%
Beginning	40	High violation	20	80.0%
High accident	40	High violation	20	73.3%
High accident	40	Beginning	40	82.5%

good records (CG), the following four variables can be used:

4. Running time. 8. Accelerator reversals. 10. Gross steering reversals. 11. Fine steering reversals.

The variables are logically divided into categories relating to (*a*) traffic, (*b*) time, (*c*) vehicle motions, and (*d*) driver-control movements.

It seems reasonable to expect that at least one variable will be needed in each category for discrimination of driver's characteristics. The final selection of five variables gives one in traffic (No. 2), one in time (No. 4), none in vehicle motions, and three in driver-control motions (Nos. 8, 10, 11).

It is reassuring to note that these five variables have about the same high confidence level when comparing any combination of drivers.

It it interesting that: (*a*) Traffic density (2) is a factor of some importance only when comparing beginning drivers with the other groups. (*b*) The vehicle-motion variables drop out, probably because speed change is related to accelerator and brake action. Thus, in this particular type of experimental design, speed change is apparently not an important parameter compared to the driver-control movements.

CLASSIFICATION OF INDIVIDUAL DRIVERS

Up to this point, for the sake of clarity, the discussion has been confined to the "discriminant" separation of drivers into groups or classes, and no attempt has been made to classify drivers on an individual basis.

But one of the primary reasons for using the multivariate-analysis method known as the Discriminate Analysis Method is that it permits an individual to be assigned to one of two populations. This follows from the fact that all X values or scores for each individual as found in analysis may be arranged in order of rank.

If two groups of drivers were entirely different, then all X scores for one group would fall in one range and all the X scores for the other group would fall in a different range. But since the groups were not entirely different there was overlapping.

The results of comparing groups by ranking are shown in Table 3.

According to the analyses just given it appears to be possible to place drivers into a

TABLE 4

DISCRIMINATORY ANALYSIS OF EXPERIENCED DRIVERS VERSUS BEGINNING DRIVERS

Classification	Control group	Beginning drivers	DFC[a]
No. tests	40	40	10^{-5}
Traffic density (2)	4555.65	5730.05	−0.9240
Running time (4)	1774.15	1806.80	+1.8246
Accelerator reversals (8)	72.87	76.37	−0.3458
Macro steering reversals (10)	565.02	644.17	−4.4884
Micro steering reversals (11)	1078.65	1189.47	−0.0515
Mean X	−.0381492	−.0191403	
Variance	$.156796 \times 10^{-3}$	$.122118 \times 10^{-3}$	
SD	±.0125132	±.01104572	

Note. N = 40 for each group. $F(5, 74) = 3.21160131$. Level of significance = 1.0% (Level of confidence = 99.0%).
[a] Discriminant function coefficient.

TABLE 5

METHOD FOR OBTAINING THE X SCORE FOR THE NEW DRIVER

Driving variable	Data	DFC[a]	
Traffic density (2)	5721	$\times (-0.9240 \times 10^{-5}) =$	$-5286.2040 \times 10^{-5}$
Running time (4)	1629	$\times (\ \ 1.8246 \times 10^{-5}) =$	2972.2734×10^{-5}
Accelerator reversals (8)	119	$\times (-0.3548 \times 10^{-5}) =$	-42.2212×10^{-5}
Macro steering reversals (10)	686	$\times (-4.4884 \times 10^{-5}) =$	$-3079.0424 \times 10^{-5}$
Micro steering reversals (11)	966	$\times (-0.0515 \times 10^{-5}) =$	-49.7490×10^{-5}
		$X =$	$-5484.9432 \times 10^{-5}$
		or	$-.05484943$

[a] Discriminant function coefficient.

preselected group at least two-thirds of the time. This was accomplished by comparing two groups. Another question is that of whether it is possible to employ the discriminant-function coefficients, established by comparing groups of drivers to individual drivers outside these groups? More specifically, for example, is it possible to select a beginning driver at random, test his driving over the prescribed test course, and determine whether his driving ability is in the beginning group or in the control group?

To illustrate the method of determining whether a preselected driver will fall into the proper category, let us select a beginning driver not previously tested and ascertain whether test drive places him in the beginning group or into the control group.

The previously established discriminant coefficients are shown in the last column of Table 4.

If the field data for the beginning driver are multiplied by the corresponding discriminant coefficients found in Table 4 and added, the X score for the new driver is obtained (Table 5).

To determine whether the score for the beginning driver (−.0548494) falls into the beginning group or not, first locate the "neutral" or midpoint between the two means. In the present example the midpoint (noting that the SD for the distribution of the experienced drivers is ± .012513 and for the inexperienced is ± .011046) is equal to:

$$\frac{(.012513)}{(.012513 + .011046)} \times -.049140 - (-.038149) + (-.038149) = -.043987.$$

Any X value less than −.043987 is in the inexperienced group and any X value greater than this is in the experienced group. The X value of −.548494 being less than −.043987 is in the inexperienced or beginning group where it should be. The farther away the X value is from the value −.043987, the more assurance that it is one group or the other.

Using this method the results shown in Table 6 were obtained.

DISCUSSION OF RESULTS

The methods and techniques described in this report add a new dimension to the understanding of driving behavior. Apparently there is revealed in ordinary driving those behavior patterns that lead to safety or probability of mishaps on the highway.

The initial series of tests showed significant differences between groups of drivers. The second series of tests, over a new course and with different drivers, has confirmed these results and indicated a high degree of confidence in the discrimination between the different groups and an identification of individual drivers to a much higher degree of reli-

TABLE 6

CLASSIFICATION OF INDIVIDUAL DRIVERS

No. preselected drivers	Preselected category	No. placed in category	Percentage correctly placed
20	Beginning	14	70%
8	High violation	8	100%
8	Low violation	6	75%
9	High accident	6	67%

ability than obtained by any other test procedure known to the authors.

Complete reliability is hardly to be expected. One should expect some overlapping in performance scores. Chance plays such an important role in a driver's record that it is always difficult to determine whether the difference between driver experiences is due to chance or whether it is due to performance. The high-accident driver may have had an abnormal degree of poor luck. With this in mind, the overlapping of the scores by about one-third is not surprising.

In setting up standard scores on a given course, for example, one could use only the lowest 50% of the scores of the high-accident group and the highest 50% of the scores in the low-accident group. If this was done, the populations, in this case, would be completely separate and nonoverlapping. Using a scale established by this method, one could test a new driver and classify him with good assurance.

It is interesting to consider why the measured variables are significant in predicting the classification of a driver. It is logical to assume that the higher the demands on a driver in performing a certain task, the greater likelihood of driver failure with attendant higher accident probability. Undoubtedly, there is some balancing effect from recognition by the driver that the more difficult the task the greater is the risk and, hence, motivation of the driver to perform at his best. But there is a limit to a particular driver's performance ability. The studies have shown that the inexperienced driver and the driver with a poor record generally makes more reversals of the controls which probably results from overcorrection and indecision. This means that these drivers are working closer to their limits than others. When an unexpected event occurs requiring an additional judgment or action by the inexperienced driver, there is less likelihood that he can cope with the situation than can the driver who is performing well within the limits of his capability.

More research is needed with the Drivometer—a device for measuring and recording driver behavior. The nature of the research clearly reveals that drivers should be studied over a long period of time. What happens to a driver's habits in 5 years? What results can be achieved by studying other groups of drivers and by refining the control group? It is necessary to know the answer to these and other questions.

It is hoped that a new avenue has been opened up to safer highways. Traveling this avenue may be long and difficult, but the rewards could be great.

REFERENCES

Bennett, C. A., & Franklin, N. L. *Statistical analysis in chemistry and the chemical industry.* New York: Wiley, 1954.

Greenshields, B. D. Driving behavior and related problems. *Highway Research Record,* Number 25. Washington, D. C.: Highway Research Board, 1963.

(Early publication received April 22, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 211–215

CAN COMPUTERS WRITE COLLEGE ADMISSIONS TESTS?

JAMES M. RICHARDS, JR.

American College Testing Program, Iowa City, Iowa

For many years psychological tests have been scored by machines, and recently computers have assembled existing items into tests and have scored essay examinations. This study goes beyond these earlier techniques and explores the possibility of computer item writing. A computer procedure for writing verbal comprehension items was developed and used to write a 72-item test. This test, together with the Wide Range Vocabulary Test, was administered to University of Iowa freshmen. The test intercorrelations, reliabilities, and correlations with grades suggest that, in principle, computers can write college admissions tests. Possible objections to computer-written tests are considered.

For many years, objective tests and machine procedures have been reducing the role of human beings in measuring the aptitude and achievement of students. The first large-scale application of machines in testing, of course, was to the scoring of tests. With subsequent advancement in the technology of scoring machines, it is now possible to score in 2 or 3 days the tests of several hundred thousand students who took a nation-wide examination. In later applications of machines, Rock (1965) has shown that computers can simulate the behavior of human test developers in assembling a test with specified properties from an item file; Page (1966) has shown that computers can score some aspects of essay examinations; and Osburn (1966) has developed a computer procedure for writing statistics problems.

Recently, large, supplementary, random-access memory devices have been developed for computers. The characteristics of such devices also suggest the possibility of writing multiple-choice test items on a computer. The purpose of the present study, therefore, was to determine whether, in principle, some tests for screening college applicants can be written by computers.

Verbal comprehension is probably the test factor most useful for predicting academic success in college. An attempt was made, therefore, to develop a procedure that a computer could use to write verbal comprehension items. Specifically, a procedure for writing synonyms items was developed. A synonyms test provides a good measure of verbal comprehension, and the use of the Wide Range Vocabulary Test (WRVT) from the Educational Testing Service Factor Kit made it possible to compare a machine-written synonyms test with a factorially pure synonyms test written by humans.

Computer Procedure

It is relatively easy to develop a procedure for choosing a stem and a correct alternative. One merely stores a dictionary of synonyms in a supplementary memory device, generates a random number, and uses this number to select a word to be the stem. One then generates another random number and uses it to select one of the synonyms for the stem as the correct alternative. If the stem word is more than one part of speech, it is necessary to choose one part of speech, and it is necessary to choose one part at random for use in the item before choosing the correct alternative.

Just this exact procedure was used to pick stem words and correct alternatives in the present study. Since no computer with a large supplementary memory device was available, the actual operation of the computer was simulated. The simulation was rigorous, however, and the items corresponded exactly to what would be written by a computer. This simulation used one of the oldest random-access supplementary memory devices—namely, a book. Although using a real computer to accomplish this task would require a considerable amount of work, the difficulties would involve the large amount of material to be stored in memory rather than how to go about using the computer. Therefore, it is the sort of task that would be assigned to a junior programmer.

After a stem and correct alternative were chosen, obviously the next step was to choose "distractor" alternatives (i.e., wrong answers). It soon became clear that developing a sensible procedure for choosing distractors is the most difficult problem in writing tests on a computer. Indeed, a perusal of the literature, after this problem became obvious, suggested that distractor alternatives are little investigated and little understood and are perhaps the most important

TABLE 1

MEAN, STANDARD DEVIATION, RELIABILITY, VALIDITY, AND INTERCORRELATION OF WIDE RANGE VOCABULARY TEST AND THE COMPUTER-WRITTEN TEST

Test	Males (N = 599)	Females (N = 613)
Wide Range Vocabulary Test		
No. items	48	48
M	23.74	25.16
SD	6.44	6.42
K-R 21 reliability	.73	.73
Correlation with first-semester GPA	.30	.30
Computer-written test		
No. items	72	72
M	53.88	55.53
SD	5.76	5.41
K-R 21 reliability	.60	.57
Correlation with first-semester GPA	.30	.32
Intercorrelation of two tests	.62	.64
Intercorrelation corrected for unreliability of both tests	.94	.98

neglected problem in the construction of multiple-choice tests. Most writers on test development merely suggest that the item writer must use his judgment and ingenuity in choosing distractors. Since computers have neither judgment nor ingenuity, such advice was useless for the present study.

The problem of distractors for synonyms items was solved in the following way. Roget's *Thesaurus* provides a classification scheme for word meanings in which each word is categorized into one or more categories. Numbers are assigned to these categories of meaning in accordance with an overall conceptual scheme. In the present study, the basic procedure for picking distractors was to choose randomly from words in adjacent categories in the *Thesaurus* scheme.

More specifically, the numerical code or, in other words, the category of meaning, shared by the stem and the correct alternative was determined. If more than one numerical code pertained, a random number was generated and used to select one of the codes. The procedure for modifying the numerical code as a step in picking distractors from adjacent categories consisted of adding 1 to or subtracting 1 from the code number for the stem and correct alternative A random number determined the choice between addition or subtraction. Then the distractor was chosen at random from words with the modified code number and of the appropriate part of speech. A check was then made to determine whether the distractor had *any* code numbers in common with either the stem or the correct alternative. If so, a new distractor was chosen at random. Additional distractors were chosen by repeating the process of modifying the code number and choosing a distractor at random. Thus, the first distractor was chosen from words one category away from the stem, the second distractor from words two categories away, etc. In a few instances, of course, all the words in a category would share a common meaning with the stem. In such cases, the procedure was simply to modify the code number and choose a distractor from the next category. The locations of the correct alternative and the various distractor alternatives in the item were also determined by random numbers.

In this way, a 72-item computer-written (or at least writable) synonyms test with four alternative items was developed. Such items and tests have several interesting properties. First of all, such items are definitely a random sample from a specified population of items. Thus, a test composed of them would conform rigorously to an assumption for several ways of estimating reliability. Second, while the population of items is very large, it is certainly finite. Similarly, while the number of alternative test forms of a given length that can be assembled from such items is very, very large, it also is finite. It appears, moreover, that the finite quality results primarily from the complete specification of the operations for writing items and may be common to many, if not most, item populations. Since the population of items and tests is very large, this has few practical implications. It may have theoretical implications, however. For example, true scores are usually defined in terms of an infinite number of tests.

Moreover, it appears that similar procedures can be used to write many other kinds of aptitude test items. With only minor modifications the procedure used in the present study could be used to write analogies items. The results of this study may be generalizable, therefore, to *some* tests of "higher" mental processes In addition, no insurmountable difficulties should be encountered in programming a computer to write mathematics items, although developing efficient distractors might be a problem.[1] The work of Osburn (1966) is a promising step toward computer-written math items.

EVALUATION OF THE TEST

The computer-written test, together with the WRVT, was administered to entering freshmen at the University of Iowa in the fall of 1965. A generous time limit permitted all students to complete the computer test. At the end of the first semester, the grade-point

[1] It is less clear that all possible types of items can be written by computers. John Holland (personal communication, 1964) has suggested a model for items in which alternatives are points equidistant in a response space. It is most unlikely that existing computers can write such items.

TABLE 2

COMPARISON OF TWO PARTS OF WIDE RANGE TEST WITH TESTS COMPOSED OF SELECTED COMPUTER-WRITTEN ITEMS

	1	2	3	4	5
Males[a]					
Wide Range Test					
1. Part 1	—				
2. Part 2	.51	—			
Computer Test					
3. Items selected for validity	.62	.43	—		
4. Items selected for homogeneity	.61	.41	.93	—	
5. Items selected for discrimination	.64	.45	.89	.88	—
K-R 21 reliability	.56	.71	.61	.60	.63
Correlation with first-semester GPA	.29	.21	.41	.35	.35
No. items	24	24	24	24	24
M	12.32	11.54	18.63	19.03	14.94
SD	3.59	4.34	3.17	3.08	3.78
Females[b]					
Wide Range Test					
1. Part 1	—				
2. Part 2	.50	—			
Computer Test					
3. Items selected for validity	.52	.56	—		
4. Items selected for homogeneity	.45	.56	.92	—	
5. Items selected for discrimination	.51	.58	.89	.88	—
K-R 21 reliability	.68	.41	.55	.53	.61
Correlation with first-semester GPA	.28	.22	.30	.29	.29
No. items	24	24	24	24	24
M	13.57	11.58	19.30	19.66	16.46
SD	4.14	3.14	2.82	2.69	3.52

Note.—Some values for males are inflated because item selection and correlations are based on the same group. Intercorrelations of computer tests are inflated by substantial item overlap.

[a] $N = 599$.

[b] $N = 613$.

average was determined for these students.[2] The means, standard deviations, Kuder-Richardson (K-R 21) reliabilities, predictive validities, and intercorrelation of the two tests[3] for 599 male and 613 female freshmen are shown in Table 1.

These results are partly bad and partly good. The items in the computer-written test, on the average, are easier than those in the WRVT (compare the mean to the length for both tests), and the reliabilities are lower for the computer test. This is disappointing because the greater length of the computer test should produce higher reliability. On the other hand, the validities are comparable to those for the WRVT, and the intercorrelation of the two tests is not too far from the limits set by reliability. On the whole, these results are encouraging enough to justify further analysis.

Since the items in the WRVT were carefully selected, while the computer test consisted of entirely unselected items, the effect of item selection should be evaluated.[4] The WRVT consists of two parts with 24 items each. Accordingly, 24-item tests were selected from the computer test by three different procedures. The item selection was based only on data for males, so that the female sample would provide a cross-validation group. The first procedure, designed to maximize validity, was simply to pick the 24 items with the highest correlation with first-semester grades.

[2] The author would like to thank Ted McCarrel, Charles B. Statler, and Willard L. Boyd for making it possible for him to obtain these data.

[3] Data analysis for this study was carried out at the University of Utah computer center.

[4] It should be noted, however, that selected items are no longer a random sample from the specified population.

The second procedure, designed to maximize homogeneity, was to pick the 24 items with the highest correlation with the computer test total. The third procedure, designed for high discrimination among students tested, was to eliminate items answered correctly by more than 90% or fewer than 10% and then choose the 24 items with the highest correlation with the computer test total. Each of the procedures, of course, could easily be incorporated in a computer program for assembling test forms.

Data for the two parts of the WRVT and the three different tests composed of selected items from the computer-written test are summarized in Table 2. It should be remembered that some values for males are inflated because the correlations and item selection are based on the same group. There is also considerable item overlap in the three computer tests, which produces much inflated intercorrelations.

While the computer test is still somewhat easier, these data confirm without qualification that synonyms tests can be successfully written on computers. Each of the three computer-written tests correlates about as highly with the two parts of the WRVT as the two parts correlate with each other; the reliability of each computer test is somewhere between the reliabilities of the two parts of the WRVT; and the predictive validities of the computer tests are, if anything, slightly higher than the validities of the WRVT.

DISCUSSION

The results clearly indicate that it is possible, in principle, for computers to write some kinds of college admissions tests. This suggests that, with a thorough application of *existing computer technology,* it would be entirely feasible, in principle, to automate many, if not all, aspects of college admissions testing. It would be possible to install in each high school an input device to a central computer at a testing agency. On this device, the computer would display test items one by one on a screen; the student would push a button to indicate his response to each item; and his response would be transmitted back to the computer. Through "time sharing," many students could be tested at once. The central computer would generate new items, include a few with each test administered, develop and equate alternate forms of the test, score each student's test, and even transmit the scores directly to another computer (or output device) at the college of the student's choice. Thus it would also be possible to eliminate answer sheets, scoring machines, and score report forms (the possibilities for computer display of items, response recording, etc., exist, of course, just as fully for tests prepared by people as they do for computer-prepared tests). All of this could take place entirely "untouched by human hands." It seems probable, moreover, that in the long run it would be to the economic advantage of test agencies to automate their operations, including item writing.

While this is possible with existing computer technology, a possibility is not an imperative, and there are many legitimate objections to such a dehumanized system. It is hoped, however, that humanists and others appalled by the *1984* overtones will make a more serious response to it than an implied Luddite computer smashing. For if computer-written college admissions tests are objectionable, rejection of them on some basis other than predictive validity, while leaving unchanged the social system making high validity possible, would be a particularly dishonest and deplorable case of removing symptoms rather than treating the underlying disease.

Any serious confrontation of the issues involved must recognize what is, on the evidence, an indisputable fact: namely, that multiple-choice tests—even those written by computers—are the best and fairest *currently available* estimates of potential for academic success in colleges as they are now constituted. This is true in the sense that an able college applicant, if he is behaving rationally in his own self-interest, will take such a test in preference to other ways of having his academic potential evaluated, such as taking an essay examination or being interviewed.

What convincing argument, then, could be advanced against having such tests written by a computer if the products should be indistinguishable from, or superior to, tests written by people? The author thinks any serious criticism of such tests must rather

begin with a criticism of conventional measures of success in college, or, to put it bluntly, of grades given by college professors. If grades can be predicted reasonably well by a completely dehumanized test, just what is wrong with grades? That something may indeed be seriously wrong is indicated by the many studies showing, at best, a negligible relationship between grades and performance outside the classroom in important areas of human endeavor (Hoyt, 1966; Richards, Holland, & Lutz, 1966).

REFERENCES

HOYT, D. P. College grades and adult accomplishment: A review of research. *Educational Record,* 1966, 47, 70–75.

OSBURN, H. G. Computer aided item sampling for achievement testing. Paper read at American Psychological Association, New York, September 1966.

PAGE, E. B. Automatic grading of student essays. In E. B. Page (Chm.), Automatic processing of natural language: New methodologies for educational research. Symposium presented at American Educational Research Association, Chicago, February 1966.

RICHARDS, J. M., JR., HOLLAND, J. L., & LUTZ, S. W. *The prediction of student acomplishment in college.* Iowa City, Iowa: American College Testing Program, 1966.

ROCK, D. Assembly of tests by use of an automated item file. *Research Memorandum 65-14.* Princeton, N. J.: Educational Testing Service, 1965.

(Early publication received November 14, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 216-222

LEVEL OF PROFICIENCY AND MULTIDIMENSIONAL VIEWPOINTS ABOUT PROBLEM SIMILARITY[1]

DANIEL LANDIS, CARL A. SILVER, JAMES M. JONES[2]

Franklin Institute Research Laboratories, Philadelphia, Pennsylvania

AND SAMUEL MESSICK

Educational Testing Service, Princeton, New Jersey

The object of this research was to determine whether individual consistencies in judgments of problem similarity are related to the proficiency level of the judges. 51 radar controllers of varying levels of experience and competency were asked to judge the degree of similarity among 30 air-traffic control stimuli using the method of successive intervals. These data were analyzed according to the "points-of-view" procedures developed by Tucker and Messick. 4 dimensions of viewpoint were extracted representing 4 major subgroups of Ss, each of which exhibited consistently different emphases in their judgment of stimulus similarity. Results indicated that (a) assignment to a viewpoint group was related to training and competency level, and (b) Ss with greater proficiency tended to view the stimuli, not in terms of their physical characteristics, but rather in terms of the responses that would be required for air-traffic control.

Several studies have indicated that not all *S*s use the same perceptual dimensions when making similarity judgments of visual stimuli. For example, Helm and Tucker (1962) found that judgments of color similarity for color-blind *S*s produced a multidimensional space quite different from that of normals. Tucker and Messick (1963) have developed analytic procedures for determining the number of different viewpoints about stimulus similarity represented in a set of individual judgments and for identifying *S*s who "perceive" the stimuli in a similar fashion. In a recent application of this technique, Silver, Landis, and Messick (1966) uncovered several dimensions of viewpoint in judgments of similarity among random geometric forms. Silver, Landis, and Jones (1965) also demonstrated the same phenomenon of consistent individual differences using visual displays resembling political situation maps. In the latter two studies no attempt was made to gather independent information on the characteristics of *S*s exhibiting each "point of view." However, Landis, Silver, and Harrison (1966) subsequently demonstrated that these viewpoints correlated with cognitive-perceptual style. The present research was designed to gather information of the relationship between these viewpoints and a training variable.

METHOD

Subjects

A total of 51 *S*s participated in this experiment, all of whom were employees of the Federal Aviation Agency (FAA) Air Traffic Control Center, Oberlin, Ohio. These 51 FAA employees represented four classifications: assistant controller, manual controller, radar controller, and coordinator. The first three groups were further subdivided into levels of competency according to the judgment of the training department of the FAA Center.

Sixteen *S*s were classified "assistant controller"; six were rated "excellent," nine "good," and one "poor." These 16 *S*s had minimal experience with interpreting radar displays. The duty of the assistant controller is to provide the more advanced controllers with flight-data strips on the outbound, inbound, and over-flying aircraft that come under the jurisdiction of their sector control.

The seven members of the second group had attained the rating of "manual controller"; two were rated "excellent," three "good," and two "poor."

[1] This research was supported by the United States Naval Training Device Center, Port Washington, New York, under contract N61339-1345, Gene Micheli, Contract Monitor. The authors wish to express their appreciation to David R. Saunders for supervising the computer analyses and John Wilson, Chief, Training Section of the FAA Oberlin Air Traffic Control Center for help in securing *S*s. Portions of this research were presented to the joint meeting of the Psychonomic Society and the Psychometric Society, Niagara Falls, Ontario, October 1964.

[2] Now at Yale University.

The duty of the manual controller is to give clearances to outbound, inbound, and over-flying aircraft as they follow flight plans within the controller's sector; he frequently uses the radar scope in his work.

Twenty-one "radar controllers" constituted the third group of Ss; eleven were judged "excellent," seven "good," and three "poor." The radar controller has the ultimate responsibility for all aircraft in his sector. He must communicate orally with the pilot, issue clearances, and observe the radar scope to be certain that all aircraft in his sector are separated. Naturally, this group and the one below had a very high degree of proficiency in the use of radar displays.

The fourth group of Ss was made up of seven coordinators, who are responsible for coordinating the entire air-traffic-control process. This group was not rated according to competency.

Stimuli

Thirty air-traffic problems were selected by the training personnel of the FAA from a larger set of such problems. The problems selected were representative of as wide a variety of such problems as possible, within the constraints imposed by the simulation equipment. The adequacy of the representation was judged by two members of the training department of the FAA Oberlin Traffic Control Center, in conjunction with air-traffic controllers familiar with the recurrent types of air-traffic problems. The traffic problems were drawn on 8 × 10-in. map overlays of the Brecksville high-altitude sector; below each map were placed flight-data strips indicating the pertinent information (speeds, distances, headings, flight plans, and aircraft identifications) for all aircraft represented in the traffic situation. Figure 1 is a sample stimulus, showing the map and flight-data strip. Each of the 435 possible pairs of traffic-control problems was constructed, and each pair was placed on a card for presentation to Ss.

Procedure

Each S was given a stack of 30 cards, each containing 1 of the 30 stimuli. Each was instructed to examine the stack, become familiar with the set, and select the two situations he judged to be least similar. The procedure of requiring each S to judge maximum dissimilarity was used to ensure some degree of familiarity with the total range of stimuli prior to the experiment proper.

Subsequently, each S was presented with the set of 435 pairs and told to divide this stack into 16 ordered categories, following a procedure previously used by Messick (1956). First, he was to divide the stack into two piles, one representing the relatively similar pairs of stimuli and the other representing the relatively dissimilar pairs of stimuli. When S had divided the cards into two stacks, he was instructed to divide each of those stacks into two in the same way. He was to continue to subdivide until there were 16 stacks ranging from the most similar

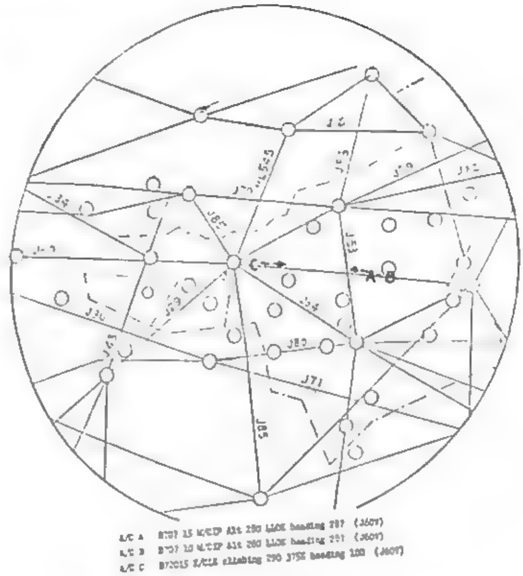


FIG. 1. Sample stimulus.

pairs to the least similar pairs. The S was given ample time to reexamine the stacks and was encouraged to make any adjustments in stimulus placement that appeared necessary. Depending upon which of the 16 categories of judged similarity a stimulus pair had been placed in, a score of 1–16 was assigned to each pair for each S.

RESULTS

The similarity scores were analyzed using the Tucker and Messick (1963) individual differences model of multidimensional scaling. This analysis first determines the number of different "points of view" about stimulus similarity (as reflected in consistent individual differences in the ratings) and then provides a separate multidimensional representation of the perceived stimulus variations according to each viewpoint. Scores representing the extent to which each individual's ratings correspond to each viewpoint are also provided.

Dimensions of Viewpoint

The individual ratings of interstimulus differences were tabulated into a matrix with one column for each S and one row for each pair of stimuli (i.e., the matrix was of order 435 × 51). A 51 × 51 matrix of sums of cross-products among these similarity ratings was then factored to determine the number of dimensions of "viewpoint" among the Ss. The pattern of characteristic roots from this analy-

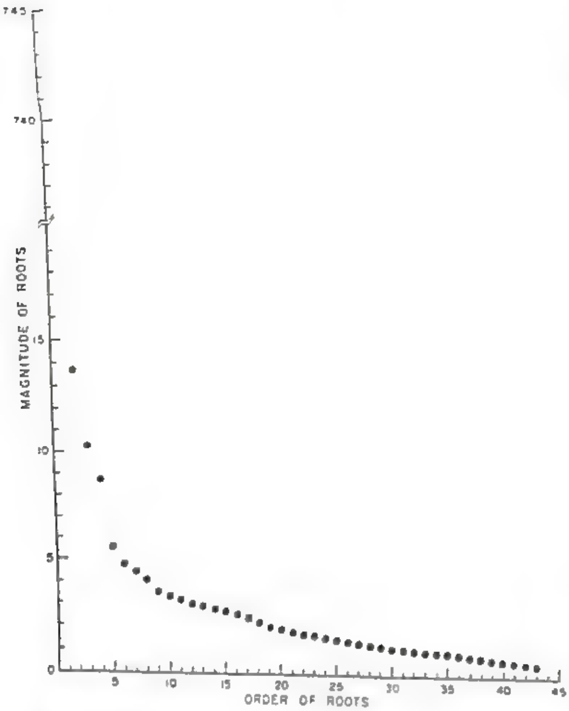


FIG. 2. Components of subject space (viewpoints dimensions).

sis (presented in Figure 2) indicates that at least four dimensions or points of view can be distinguished among the *S*s. The four factors corresponding to the four largest latent roots were rotated using the Equamax approximation to simple structure (Saunders, 1962). A factor loading was thus obtained for each *S* on each rotated factor or viewpoint dimension (in this case the factor matrix was of order 51×4). The size of these factor loadings indicates the extent to which each *S*'s ratings corresponded to the particular viewpoint.

In order to see if these points of view about stimulus similarity were in any way related to individual levels of training and competence, *S*s were classified into four subgroups representing the major viewpoint dimensions; this classification was then compared with the classification based upon training and competence. The rationale for the subgroups was as follows: Each *S* receives a factor loading on each of the four viewpoint dimensions. According to the individual differences model of multidimensional scaling, the size of these loadings indicates the weight each dimension receives in reproducing *S*'s individual similarity ratings by a linear combination of the four factors. The dimension on which an *S* receives his highest loading thus represents the dominant influence in his own individual view of the stimulus characteristics. For the present purposes, *S*s were classified in a particular viewpoint group if they had the same dominant dimension (i.e., if they received their highest loading on the same viewpoint factor). There were thus four subgroups, one corresponding to each of the viewpoint dimensions. However, even though the dimensions had been rotated in this study to capitalize upon whatever simple structure occurred in the sample, most *S*s tended to be distributed throughout the four-dimensional space, having moderately high loadings on more than one dimension. It was recognized that the classification of such *S*s by the present rule would tend to dilute the homogeneity of viewpoint within each group, but it was hoped that the simplicity of the assignment would permit a quick appraisal of the relation between viewpoint and competence that could always be elaborated by more refined categorizations.

The relation of training and competence levels to viewpoint placement was striking. Viewpoint Group C clearly represented a skilled group; 11 of the 13 *S*s assigned to this group had been classified either "radar controller" or "cooordinator." Further, it is interesting to note that of the 2 *S*s that do not seem to belong in Viewpoint C because of their low experience classification, 1 was considered by his superiors to be an extremely capable controller who would move up the ranks quickly.

Similarly, Viewpoint Group A seemed to be made up primarily of unskilled personnel. Ten of the 16 *S*s assigned to this viewpoint were at the assistant-controller level. Viewpoints B and D were not as clearly drawn as the other two, but there were attributes of their associated stimulus spaces (discussed below) that appear to differentiate them, with D appearing to represent a somewhat higher level of competence than B.

Stimulus Dimensions

For each viewpoint, interstimulus distances were then estimated by averaging the similarity ratings for all *S*s in that viewpoint

group. In the language of the Tucker-Messick (1963) model, distance estimates were obtained for an "idealized individual" placed at the average or centroid of the real individuals in each group. The four resulting distance matrices were then analyzed by standard multidimensional scaling procedures (Messick & Abelson, 1956) to obtain the dimensions of perceived stimulus variation according to each viewpoint. The characteristic roots from these four multidimensional scaling analyses are portrayed in Figure 3. The pattern of these roots indicates that stimulus similarity according to Viewpoint Group A involves several dimensions of stimulus variation—at least three large dimensions and perhaps as many as eight altogether. Viewpoint Group B, on the other hand, appears to emphasize only about two dimensions. Viewpoint C has one large dimension and perhaps three moderate-sized ones, while Viewpoint D appears to stress only two dimensions, with one much larger than the other. For each group separately, the obtained stimulus factors were rotated analytically to the Equamax criterion of orthogonal simple structure (Saunders, 1962).

In order to identify or "name" the stimulus dimensions thus obtained, the following procedures were adopted: The stimuli were ranked according to their factor loadings on each dimension. Skilled FAA personnel (Air Traffic Control instructors) were asked to view these ranked stimuli and to suggest names for the various dimensions. The names obtained were then used as a basis for developing a set of hypotheses about the underlying stimulus continua. Six hypotheses were derived in this manner:

1. Amount of watching.
2. Number of aircraft.
3. Number of directions.
4. Amount of analysis required.
5. Severity of problem.
6. Necessity for action.

Hypotheses 2 and 3 refer to stimulus characteristics and lead directly to a numerical index for each stimulus. The other hypotheses refer to characteristics of responses required for air-traffic control. Quantitative indexes for these four hypotheses were obtained by rank-

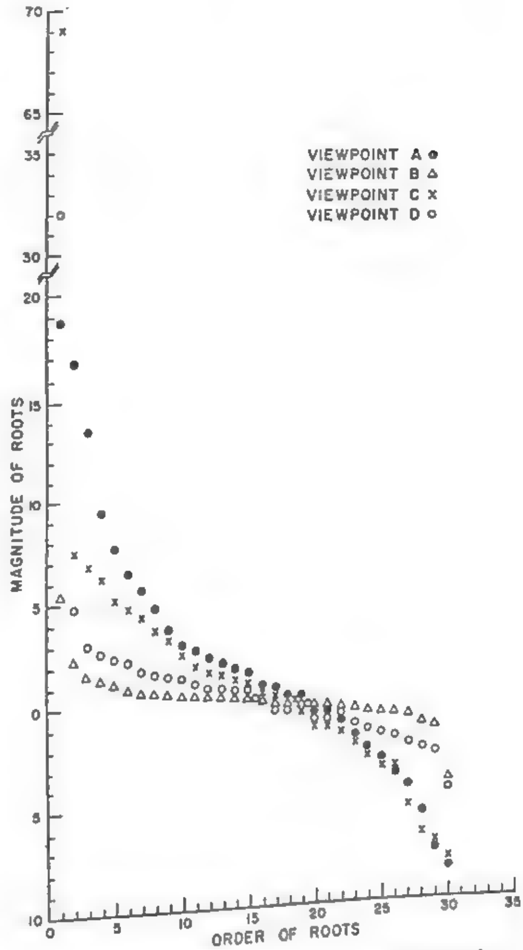


FIG. 3. Components of stimulus space (stimulus dimensions).

ing each stimulus according to the following considerations:

"Watching" (in Hypothesis 1) was defined as the degree of continuous monitoring of the radar scope required to ensure that two or more aircraft on possibly conflicting courses will pass with the required separation. Sometime aircraft are not on conflicting courses, but their present separation is marginal and any unforeseen circumstances (such as pilot lapse, weather, or technical communication difficulties) could produce a conflict situation.

"Amount of analysis" (Hypothesis 4) was derived from the number of computations necessary to determine that an adequate separation is being maintained for all aircraft in the sector.

"Severity" (Hypothesis 5) refers to the complexity of a given situation. It is a com-

TABLE 1

RANK CORRELATIONS BETWEEN STIMULUS LOADINGS ON EACH DIMENSION AND STIMULUS VALUES ACCORDING TO SIX HYPOTHESES

Hypothesis	Correlational value							
	Viewpoint A	Viewpoint B		Viewpoint C			Viewpoint D	
	D-6	D-1	D-2	D-1	D-2	D-3	D-1	D-2
Watching	−.08	−.14	.37*	.55**	.47*	.55**	.62***	−.26
No. aircraft	.22	−.11	.03	−.03	.09	.18	.01	−.12
Direction	−.38*	−.30	.22	.72***	.55**	.33*	.34*	−.19
Amount of analysis	.08	−.54**	.40*	.43*	.37	.53*	.55**	−.16
Severity	−.14	−.12	.35*	.65***	.49**	.55**	.51**	−.27
Action	−.14	−.04	.53**	.55**	.63***	.58***	.41*	−.26

Note.—Abbreviated: D = Dimension.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

posite of the amount of watching, analysis, and action. Situations that involve a great deal of action are considered severe. If two situations require immediate action and one also requires considerable analysis, the latter is judged more severe.

"Action" (Hypothesis 6) refers to the immediacy of controller action required in a given situation. Stimuli are ranked high if action is required upon an immediate inspection of the figure (for example, when two aircraft at identical altitudes are converging at the rate of 1,000 knots air speed) and intermediate if action is determined only after some degree of analysis.

Because of the complex nature of Hypotheses 4, 5, and 6, it was not possible to derive objective indexing rules. Instead, stimuli were subjectively ranked for each hypothesis by independent judges using all the information contained in the protocols; interrater reliability was greater than +.90.

The stimulus ranks based on each of the six hypotheses were then correlated with the stimulus ranks given by the factor loadings on each dimension. These correlations (for those dimensions yielding significant values) are presented in Table 1. For Viewpoint A, there were no significant correlations for five of the dimensions, and only one significant correlation on Dimension 6 (with Hypothesis 3). The failure to develop any hypotheses that would be related to the several dimensions of Viewpoint A may have resulted from the particular manner in which these hypotheses were constructed. It will be recalled that names for the various dimensions were solicited from radar-training personnel, and it may be that these highly experienced individuals were not sensitive to the kinds of dimensions typically used by less experienced *Ss*. It would probably have been a better procedure to have the less proficient *Ss* attempt to name the dimensions as well.

Upon examination, the dimensions of Viewpoint A appeared to reflect similarities in stimulus characteristics rather than similarities in the kinds of responses required in the situation (as embodied in Hypotheses 1, 4, 5, and 6). The stimulus situations on Dimension 1 range from those involving level-flying aircraft heading in one direction with no convergences or conflicts, to those depicting one aircraft descending to land with no other aircraft in the area. The stimuli on Dimension 2 range from those situations in which aircraft are converging and changing altitudes, but with no conflict, to those in which aircraft are flying in one direction with no convergences and no conflicts. The third dimension reflects the degree to which situations depict weather problems. The basis of the fourth dimension appears to be a degree of conflict, but this dimension is small and ac-

counts for relatively little variance. At one extreme of this dimension are aircraft flying in one direction overtaking and descending through the altitude of preceding aircraft, and at the other extreme are two aircraft at the same altitude closing head-on at a very fast rate.

Examination of the stimulus dimensions for the skilled-operator group (Viewpoint C) reveals a striking difference between them and the unskilled group (Viewpoint A): all of the hypotheses except "number of aircraft" correlate significantly with factor loadings on the major dimensions. Since four of the five significant hypotheses (all except "number of flight directions") represent response requirements, it seems reasonable to conclude that skilled radar controllers differentiate among air-traffic control problems more in terms of varying operational responses they require than in terms of the stimulus characteristics per se. The stimulus situations on Dimension 1 of Viewpoint C range from those that require no immediate action but necessitate prolonged monitoring (particularly to avert any possible conflict because of weather), to those that require not only careful watching but much analysis and usually considerable action. Stimuli on Dimension 2 range from those involving unidirectional flight, no convergences, little analysis, and a fair amount of watching, to those requiring considerable watching, analysis, and action.

As might be expected, Viewpoints B and D, which characterize *Ss* from intermediate proficiency classifications, contain dimensions that are partially representative of both Viewpoint A and Viewpoint C. One dimension in both Viewpoints B and D (B-2, D-1) is similar to those found in Viewpoint C, in that it correlates significantly with the four hypotheses based upon response requirements. This dimension may thus represent the response-determined components of similarity judgments for those two groups. In addition, however, Viewpoints B and D both contain another dimension that is not related to response requirements; this other dimension, particularly in Viewpoint B, appears to be determined primarily by stimulus characteristics (see Table 1). Since the response-related dimension in Viewpoint D is larger and accounts for more of the interpoint distance variance than its counterpart in Viewpoint B, it would seem that the individuals in Viewpoint Group D tend to judge problem similarity more in terms of response requirements than do individuals in Viewpoint Group B. Furthermore, since response-related dimensions are characteristic of the skilled approach of the most proficient group (Viewpoint C), it would appear that Viewpoint Group D more closely approximates this skilled approach than does Viewpoint Group B and that they might thereby be considered to have attained a more advanced level of proficiency than that achieved by Group B.

Discussion

The results of this investigation indicate that radar controllers exhibit consistent individual differences in their viewpoints about the similarity of air-traffic control problems and that these viewpoints are related to individual levels of training and competence. The judgments of the least skilled group (Viewpoint A) appeared to be based upon many simple dimensions depicting various stimulus characteristics, whereas the judgments of the most skilled group (Viewpoint C) appeared to be based upon a small number of complex dimensions reflecting the response requirements of the stimulus situations. Viewpoint groups of intermediate proficiency appeared to utilize one complex dimension of response requirements and one simpler dimension of stimulus characteristics. The overall pattern of this relation between proficiency level and judgmental consistencies suggests that different approaches to the training of radar controllers might be profitably investigated—in particular, one in which training is initiated by considering from the outset the various relational configurations that occur in traffic-control situations along with their attendant response requirements, rather than by first considering simple, clear-cut displays and their associated stimulus characteristics.

Landis et al. (1966) have demonstrated a relationship between points of view and cognitive-perceptual style using measures that presumably are not variable over time (Gardner,

Holzman, Klein, Linton, & Spence, 1959). A longitudinal study investigating the relationship between cognitive-perceptual style and proficiency level might very well indicate whether or not proficient controllers do, in fact, change their ways of perceiving or do they become proficient because they come into the program with certain styles of perceiving and thinking. Such a study would have important implications for training programs.

REFERENCES

GARDNER, R., HOLZMAN, P. S., KLEIN, G. S., LINTON, H., & SPENCE, D. P. Cognitive control: A study of individual consistencies in cognitive behavior. *Psychological Issues*, 1959, 4, 1–185.

HELM, C. E., & TUCKER, L. R. Individual differences in the structure of color perception. *American Journal of Psychology*, 1962, **75**, 437–444.

LANDIS, D., SILVER, C. A., & HARRISON, S. Relationship between multidimensional points-of-view and cognitive-perceptual style. Paper presented to Midwestern Psychological Association, Chicago, May 1966.

MESSICK, S. An empirical evaluation of multidimensional successive intervals. *Psychometrika*, 1956, **21**, 367–375.

MESSICK, S., & ABELSON, R. P. The additive constant problem in multidimensional scaling. *Psychometrika*, 1956, **21**, 1–16.

SAUNDERS, D. R. Trans-varimax: Some properties of the ratiomax and equamax criteria for blind orthogonal rotation. *American Psychologist*, 1962, **17**, 395–396. (Abstract)

SILVER, C. A., LANDIS, D., & JONES, J. M. Development of criteria for evaluation of large-screen displays. Rome Air Development Center Technical Report 65-198, 1965.

SILVER, C. A., LANDIS, D., & MESSICK, S. Multidimensional analysis of visual form: An analysis of individual differences. *American Journal of Psychology*, 1966, **79**, 62–72.

TUCKER, L. R., & MESSICK, S. An individual differences model for multidimensional scaling. *Psychometrika*, 1963, **38**, 333–367.

(Early publication received December 31, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 223–225

VOCATIONAL INTERESTS AND ACCIDENT PRONENESS [1]

JOSEPH T. KUNCE

University of Washington

The interrelationships of an index of accident proneness (AP) obtained from the SVIB with accident rate, age, job tenure, and job hazard were investigated. A sample of 62 industrial employees including unskilled and professional workers was studied. The following results were obtained: a higher than average accident rate was significantly related to high job hazard (phi = +.31) and high AP scores (phi = +.28). A longer than average job tenure was significantly related to a low accident rate (phi = −.38) and current placement in less hazardous jobs (phi = −.27). Ss having long tenure also had significantly lower AP scores (phi = −.29), and older Ss also tended to have lower AP scores.

This report is fifth in a series of investigations (Kunce & Brewer, 1966; Kunce, Sturman, Longhofer, & Castor, 1966, Kunce & Worley, 1966, 1967) studying the relationship of psychological factors in accidents. The Strong Vocational Interest Blank (SVIB) was utilized in each study to test the hypothesis that a person's style of life (reflected in his interests) is related to incurrence of accidents. Research elsewhere has demonstrated that the SVIB measures relatively enduring interest patterns, therefore showing considerable potential for yielding an index of life style which could be evaluated with respect to accident-engendering behavior. Two scales were specifically selected to give indexes of "adventuresomeness" (Aviator scale) and "cautiousness" (Banker scale). An index of accident proneness (AP) obtained by subtracting the banker standard score from the aviator standard score was found to be significantly related to:

1. Incurrence of *severe* physical disabilities in a group of 84 adult male patients.
2. School records of accidents over a 3-yr. period for 146 high school students *prior* to testing.
3. Visits to a student health clinic for minor injuries (but not for illness) over a 2-yr. period for 112 university students *following* testing.
4. Judgments of "accident prone behavior" of 116 psychiatric patients.

In each of these studies, *S*s having AP scores of 11 or more had approximately twice the number of accidents than *S*s having lower scores.

The present study was designed to further test the validity of AP scores as a measure of accident proneness for a population of industrial workers. In addition, this type of population enabled us to investigate the interaction of AP scores with other relevant variables such as employment stability and job hazard.

METHOD

Subjects. The *S*s were 62 male employees of Birds Eye Manufacturing Plant, a division of General Foods Corporation, located in Washington state. A wide variety of employees were asked whether they would voluntarily participate in a research project. Those tested were members of the regular work force covering a broad range of occupations including supervisory personnel, engineers, accountants, electricians, mechanics, and production workers. Their ages ranged from 23 to 63 with a mean age of 41 yr. Years of employment ranged from 3 mo. to 23 yr. with a mean of 9 yr.

Instruments. The SVIB was given to each *S* and a score for accident proneness (AP) was obtained as previously described by subtracting the standard score for the Banker scale from that of the Aviator scale. From the findings of the previous studies, AP scores greater than 10 were considered "high," +10 through −10, "average," and less than −10, "low."

An index of accident behavior was determined by dividing the total number of visits for accidents to the industrial nurse by the total number of years worked. The rates so obtained ranged from 0.00 to 8.00 per year with an average of 1.81.

[1] This research was supported in part by Public Health Service Grant CH 51-02-C-65 and Vocational Rehabilitation Grant RT-3. In addition, the studies could not have been carried out without the interest and cooperation of Clarence Baker, personnel manager, General Foods Corporation, Walla Walla, Washington.

TABLE 1

INTERRELATIONSHIP OF ACCIDENTS, AGE, YEARS EMPLOYED, JOB HAZARD, AND AP SCORES

Variable	2 Age	3 Years employed	4 Job hazards	5 AP scores	*M*	*SD*
1. Accident rate	−.02	−.38**	+.31*	+.28*	1.81	1.98
2. Age		+.08	+.18	−.22	41.45	11.01
3. Years employed			−.27*	−.29*	9.04	5.53
4. Job hazard				−.04	3.85	1.34
5. AP scores					11.02	15.50

* $p < .05$.
** $p < .01$.

Estimates of job hazard for each job were made by the plant safety officer using a 6-point scale of 1 (minimal) to 6 (maximal). The distribution of *Ss* in Categories 1 through 6 were 4, 5, 17, 11, 20, and 5. Examples of job types in each category, respectively, were accountant, stock clerk, production supervisor, lift truck operator, maintenance mechanic, and viner mechanic.

Procedure. The interrelationships of AP scores, accident rate, job hazard, age, and years of employment were evaluated by computing phi coefficients. This method was used in preference to product-moment correlation, since the data, particularly AP scores and accident rates, were not distributed normally. To make the necessary computations, the age, accident-rate, and duration-of-employment variables were dichotomized at the mean. The job-hazard variable was dichotomized into nonhazardous (ratings of 1, 2, or 3) and hazardous (4, 5, or 6). AP scores were dichotomized as "normal" (10 or less) and "high" (11 or higher).

RESULTS

A higher than average accident rate was significantly related to high job hazard (phi = +.31) and high AP scores (phi = +.28). A longer than average job tenure was significantly related to a low accident rate (phi = −.38) and current placement in less hazardous jobs (phi = −.27). The *Ss* having long tenure also had significantly lower AP scores (phi = −.29), and older *Ss* also tended to have lower AP scores. (See Table 1.)

DISCUSSION

Tables 2 and 3 were constructed to illustrate the personal factor in accidents, taking into consideration the interaction of job hazard with AP scores and job tenure with AP scores.

In the hazardous jobs, 64% of the accident-prone *Ss* had a higher than average accident rate in contrast to 39% of the less-accident-prone *Ss*. In nonhazardous jobs the percentages were 36 and 0, respectively. A similar pattern is obtained with respect to the interaction of job tenure with AP scores and accident rate.

Further investigation of the data for employees carrying out jobs rated 6, the most hazardous, revealed that only 1 of the 5 employees in that group was "accident prone" having an AP score greater than 10. Furthermore, this employee was the only one of the

TABLE 2

PERCENTAGE OF EMPLOYEES CATEGORIZED BY AP SCORES AND JOB HAZARD HAVING A HIGHER THAN AVERAGE ACCIDENT RATE

AP scores	Job hazard: Nonhazardous	Job hazard: Hazardous
High	36%	64%
Normal	0%	39%

TABLE 3

PERCENTAGE OF EMPLOYEES CATEGORIZED BY AP SCORES AND JOB TENURE HAVING A HIGHER THAN AVERAGE ACCIDENT RATE

AP scores	Job tenure: Below average	Job tenure: Above average
High	61%	22%
Normal	38%	12%

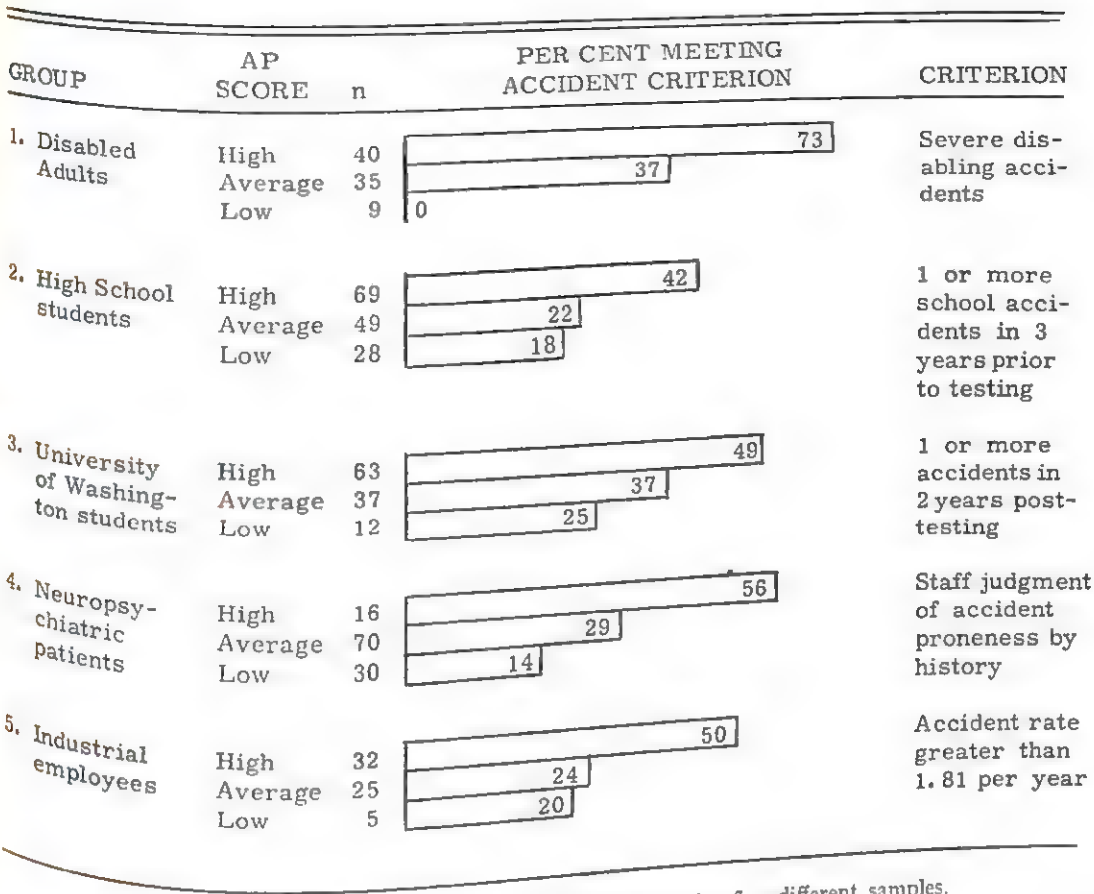


FIG. 1. Summary of AP scores and accident proneness for five different samples.

five having an accident rate greater than the average of 1.81 for the total sample.

A summary of the results obtained for five quite different populations is given in Figure 1. In each of these groups, the percentage of *Ss* meeting the accident criterion is markedly higher for the high AP group.

CONCLUSION

The successful performance on any job is, of course, dependent upon a myriad of personal and physical characteristics; hence, generalization of these data to job selection and placement would be premature and would *not* be warranted. However, these data do provide a means for identifying and classifying individuals with certain predictable characteristics which could be useful in further research. For example, do typical safety programs have the same relative impact upon high- and low-accident-prone individuals?

REFERENCES

KUNCE, J., & BREWER, B. Neuropsychiatric patients, accident proneness, and interest patterns. *Journal of Psychology*, 1966, **63**, 287–290.

KUNCE, J., STURMAN, J., LONGHOFER, P., & CASTOR, M. Psychological factors in high school accidents. *Personnel & Guidance Journal*, 1966, **45**, 140–143.

KUNCE, J., & WORLEY, B. Interest patterns, accidents, and disability. *Journal of Clinical Psychology*, 1966, **22**, 105–107.

KUNCE, J., & WORLEY, B. Interest scores and accidents among students. *Journal of College Student Personnel*, 1967, **8**(2), 121–123.

(Early publication received January 13, 1967)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 226–232

CONTRIBUTIONS OF PROJECTIVE TECHNIQUES TO ASSESSMENT OF MANAGEMENT POTENTIAL

DONALD L. GRANT
American Telephone and Telegraph Company

WALTER KATKOVSKY
Fordham University

AND DOUGLAS W. BRAY
American Telephone and Telegraph Company

The contributions of projective techniques to assessment-center staff evaluations and the relationships of projective variables to progress in management are presented. The projective data were obtained by coding reports written by a clinical psychologist from 3 projective instruments. Analyses of the data show that the projective reports particularly influenced the assessment staff in rating such characteristics as work motivation, passivity, and dependency. In addition, several of the projective variables are reliably related to progress in management, especially those pertaining to leadership and achievement motivation. In brief, the findings clearly indicate that relevant information on managerial motivation was obtained from the projective reports.

An assessment center for the measurement of potential for business management has recently been described and analyzed (Bray & Grant, 1966). Among the analyses presented are the contributions of the various assessment techniques to staff judgments and the relationship of each assessment technique, as well as the staff judgments, to a salary criterion. The assessment center studied is part of the Bell System Management Progress Study (Bray, 1964), a longitudinal investigation of the development of young men in a business management environment.

Two of the assessment techniques used were not included in the previous analysis, that is, the interview and the projective techniques. This article presents findings on the projectives; a similar study of the interview is in process and will be reported later.

Although projective techniques have been rather widely applied to the evaluation of managers, their usefulness has not been convincingly demonstrated (Kinslinger, 1966). Neither has their contribution to the judgments of assessment-center staffs been analyzed even though projectives have been used in most of the major trials of the assessment process. Where data have been reported they have focused on the relationships of evaluations based on the projective techniques to criterion performance (Hague, Otis, & Prien, 1962; Huse, 1962; Kelly & Fiske, 1951; MacKinnon, 1958). In general, the reported correlations have been low and, frequently, not reliably greater than zero.

Practically no information, however, on the contributions of the projective techniques to the evaluations of the assessors has been reported. The omission apparently has reflected a general lack of concern with the role separate techniques play in the total assessment process.

The Management Progress Study data afford an opportunity to examine the contribution of certain projective methods to assessment-center judgments, which have been shown to be predictive of progress in management, and the relationships of projective variables directly to a progress criterion. Neither the projective data nor any other assessment information have been available to Bell System management or the subjects (*S*s). There is, in other words, no contamination of the progress criterion.

Projective Tests

Three projective techniques have been employed in the assessment centers of the Management Progress Study. These are: (1) the Rotter Incomplete Sentences Blank, (2) the Management Incomplete Sentences Test,[1] and (3) six cards from the Thematic Apperception

[1] This test was constructed by Walter Katkovsky, Vaughn C. Crandall, and Julian B. Rotter.

Test (Cards 2, 6BM, 7BM, 8BM, 14, and 16). The Management Incomplete Sentences Test was constructed specifically for the study and consists of 57 items. The sentence stems were selected to assist in assessing a wide range of variables pertaining to management attitudes and performance such as achievement orientation, level of aspiration, job involvement, attitudes toward different types of assignments, and effective reactions to superiors, peers, and subordinates.

The TAT was administered in small groups of from three to six *S*s; each *S* was presented with each picture individually. The *S*s were requested to write their stories using the following three questions as a guide: What is happening in the picture? What occurred previously which brought about the circumstances you see in the picture? What will happen in the future; how will the story turn out?

The projective tests were administered in small groups early in the assessment program after *S*s had received a general orientation concerning the purposes of the program and the confidentiality of the information obtained about them. The protocols then were evaluated by a clinical psychologist who wrote one report on the impressions and hypotheses he formulated about each individual based on all three instruments. Two clinical psychologists composed the 355 reports.[2] Each presented a general description of the *S*'s personality with emphasis given to characteristics which might influence his management career, such as his general level of adjustment and specific attitudes and feelings about his job, career, family, and self. The intent was to follow an ideographic approach in these reports rather than attempt to fit the individual case to a prescribed list of variables. As the program progressed, however, a large degree of consistency did occur in the dimensions cited in individual reports, although there was considerable variability in the elaboration given to any one personality dimension.

The projective test report on a given *S*, together with reports written by other staff members based on different assessment techniques, was read to the assessment staff at the time that *S* was being considered. The staff, consisting of nine assessors, used all of the impressions and information presented from the various techniques administered during the immediately preceding 3½-day period in making their ratings on the Management Progress Study variables (see Bray & Grant, 1966, for details of the entire process).

METHODS OF ANALYSIS

For purposes of statistical analysis the projective test reports on 355 participants in the study were coded independently by two psychologists who had not participated in the assessment. Using a manual the coders rated each report on nine variables. These variables were selected because of the frequency with which they were used in the reports and were defined in terms of the descriptive phrases used by the psychologist in his reports. The variables and their definitions were:

Achievement Motivation
How ambitious, motivated, and interested in advancement and success is he?

Self-Confidence
How confident is he that he will succeed in his work? To what extent does he exhibit self-doubts and anxieties concerning success?

Work or Career Orientation
How important is work in this man's life compared with other things, such as family and recreation? How much satisfaction does he get from his job compared with other things in life?

Dependence
To what extent is he described as needing or seeking help, advice, direction, and encouragement from others?

Affiliation
How interested is he in being liked and accepted by others, being a part of groups, helping others, and avoiding arguments and friction with others? To what extent does he participate in group situations and how outgoing is he?

Optimism-Pessimism
How optimistic or pessimistic is he about life in general? How positively or negatively does he react to most of his experiences? How satisfied or dissatisfied is he in general?

Willingness to Assume a Leadership Role
How readily will he make decisions and accept leadership or supervisory responsibility? How strong are his dominance or leadership needs?

Willingness to Accept a Subordinate Role
How willing is he to act as a follower or subordinate in his relationships with others? How suggestible and submissive is he?

[2] Approximately 80% of the projective reports were composed by Walter Katkovsky and the remainder by Joseph Rychlak.

TABLE 1

Coder Reliabilities

Projective variable	College sample (N = 207)		Noncollege sample (N = 148)	
	r_{ij}	r_{ii}	r_{ij}	r_{ii}
Optimism—Pessimism	.77	.87	.77	.87
General Adjustment	.74	.85	.80	.89
Self-Confidence	.82	.90	.79	.88
Affiliation	.89	.94	.84	.91
Work or Career Orientation	.84	.91	.84	.91
Leadership Role	.84	.91	.86	.92
Dependence	.85	.92	.83	.91
Subordinate Role	.81	.90	.82	.90
Achievement Motivation	.87	.93	.85	.92

General Adjustment

> To what extent does he indicate good adjustment (high self-confidence, ability to adapt to difficult situations and accept frustrations) or poor adjustment (low self-confidence, insecurity, stress or anxiety reactions, emotional conflict, and problems)?

A 5-point scale was used in rating each variable. As an example the scale for Achievement Motivation is shown below:

> 1—Described as "very low," "well below average for the group," or "lacking" in ambition and motivation for advancement. Getting ahead is not important to him.
>
> 2—Described as "somewhat below average," "on the low side," or "low average" in ambition and motivation. Success in his work is not as important to him as other things in his life.
>
> 3—Described as "average" in ambition and motivation. He wants to advance, is moderately ambitious, but is not extreme. Success in his work shares equal importance to him to such things as achieving happiness and security.
>
> 4—Described as "above average," "on the high side," "moderately high," or as having "fairly strong" ambition and motivation. Advancement, monetary gain, or status are important to him, and he gives a fair amount of thought to getting ahead.
>
> 5—Described as "very ambitious" or "determined to get ahead" or as "very high" or "well above average." Achieving success in his work is one of the most important things to him.

The first step in analyzing the data was to group the coded reports according to the educational backgrounds of the participants. One group consisted of men who had graduated from college prior to employment ($N = 207$), the other of the men who had not attended college prior to employment but who had been promoted to management early in their careers ($N = 148$). The great majority of the men were in their 20s when assessed; a few of the noncollege men were in their early 30s. The *Ss* came from several regions of the country.

The reliability of the coding procedure was determined by correlating the ratings of the two

TABLE 2

Intercorrelations of Projective Variables

Projective variable		1	2	3	4	5	6	7	8	9
College sample (N = 207)										
Optimism—Pessimism	(1)	—	73	58	23	28	20	−08	06	14
General Adjustment	(2)	73	—	61	09	35	40	−22	−03	24
Self-Confidence	(3)	58	61	—	−15	30	53	−37	−23	45
Affiliation	(4)	23	09	−15	—	13	−36	43	47	−31
Work or Career Orientation	(5)	28	35	30	13	—	35	−03	−02	54
Leadership Role	(6)	20	40	53	−36	35	—	−57	−61	56
Dependence	(7)	−08	−22	−37	43	−03	−57	—	66	−38
Subordinate Role	(8)	06	−03	−23	47	−02	−61	66	—	−41
Achievement Motivation	(9)	14	24	45	−31	54	56	−38	−41	—
Noncollege sample (N = 148)										
Optimism—Pessimism	(1)	—	74	61	29	06	33	−26	−08	00
General Adjustment	(2)	74	—	71	29	16	45	−39	−11	13
Self-Confidence	(3)	61	71	—	−00	15	57	−57	−37	37
Affiliation	(4)	29	29	−00	—	07	−17	23	36	−28
Work or Career Orientation	(5)	06	16	15	07	—	35	−16	−15	50
Leadership Role	(6)	33	45	57	−17	35	—	−68	−62	61
Dependence	(7)	−26	−39	−57	23	−16	−68	—	66	−44
Subordinate Role	(8)	−08	−11	−37	36	−15	−62	66	—	−49
Achievement Motivation	(9)	06	13	37	−28	50	61	−44	−49	—

coders on each variable for each group of participants. The Spearman-Brown prophecy formula then was applied to the correlations.

The ratings of the coders were pooled, and the resulting sum scores of the projective variables intercorrelated and correlated with judgments of the assessment staff, scores from other assessment techniques, and progress in management as reflected by salary progress. In essence, the analyses made are comparable to those for the other assessment techniques previously studied (Bray & Grant, 1966).

Results

The reliability of the coding process is shown in Table 1, which presents the correlations between the ratings and the estimated reliabilities of the combined ratings. The latter range from .85 to .94 with a median of .91. Though the estimated reliabilities of the combined ratings vary somewhat from variable to variable, they are in general satisfactorily high.

The intercorrelations between the variables, shown in Table 2, provide useful information. The data indicate that the projective reports as coded did discriminate between the variables. Though the matrices have not been factor analyzed, the variations in magnitude of the correlations are sufficiently great to suggest that the "halo" across variables is relatively small.

Furthermore, several meaningful clusters emerge. Two major groupings can be observed in both matrices. An "adjustment" cluster is indicated by the high correlations between Optimism-Pessimism, General Adjustment, and Self-Confidence. A "leadership" or "dominance" cluster is indicated by the overlap between Leadership Role, Dependence, and Subordinate Role. Achievement Motivation overlaps the latter cluster and also has a relatively high correlation with Work or Career Orientation. Affiliation has relatively low correlations with the other variables, particularly in the noncollege sample, though it overlaps to some extent with Dependence and Subordinate Role in the college sample. Self-Confidence correlates almost as highly with Leadership Role (and Dependence, negatively, in the noncollege sample) as it does with Optimism-Pessimism and General Adjustment.

The correlations between each of the projective-report variables and variables reflecting the evaluations of the assessment staff are shown in Table 3. The latter represent scores based on factorial analyses of 25 characteristics rated by the staff (Bray & Grant, 1966). The characteristics were selected for their relevance to the study and include managerial skills, interpersonal relationships, general abilities, motives, values, and attitudes. The factors obtained and their designations are as follows:

Factor	Identification
I	General effectiveness
II (college sample only)	General effectiveness
III (college sample only)	Passive dependency
IV	Administrative skills
V	Interpersonal skills
VI	Control of feelings
VII	Intellectual ability
VIII	Work-oriented motivation
IX	Passivity
X	Dependency
XI (college sample only)	Nonconformity

In addition to the relationships of the projective-report variables to scores based on the factors, Table 3 shows their correlations with staff predictions of the likelihood of progress in management.

In general, the projective variables reflecting "leadership" motivation tend to correlate much higher with the staff judgments than do those reflecting "adjustment." Furthermore, the projective variables, as might be expected, correlate higher with judged motivational characteristics, that is, work-oriented motivation, passivity, and dependency, than they do with such ability characteristics as administrative skills, interpersonal skills, and intellectual ability. It would appear that the projective reports had a marked influence on the judgments of the assessment staffs in evaluating the motivational characteristics of the participants in the study.

That aspects of the projective reports also influenced the overall judgments of the assessment staffs can be seen from the relatively high correlations of several projective variables with general effectiveness (Factor I) and the staff prediction (will make middle management in 10 or less years). Those for Leadership Role, Dependence, and Achievement Motivation tend to be the highest. It also can be noted that while the correlations

TABLE 3

CORRELATIONS WITH STAFF JUDGMENTS

Projective variable	Staff prediction	Factor I	II	III	IV	V	VI	VII	VIII	IX	X	XI
College sample (N – 207)												
Optimism—Pessimism	11	13	14	−08	08	17	19	−05	16	−10	−06	−03
General Adjustment	10	25	23	−22	23	21	28	−01	23	−18	−19	11
Self-Confidence	24	31	21	−42	21	17	24	01	15	−39	−33	22
Affiliation	−07	−16	01	44	−08	07	−01	−08	13	26	46	−21
Work or Career Orientation	21	27	32	−05	29	17	13	10	50	−14	03	−01
Leadership Role	35	48	34	−56	36	20	29	22	25	−47	−49	36
Dependence	−30	−44	−29	55	−30	−16	−28	−33	−05	39	49	−54
Subordinate Role	−25	−32	−17	46	−24	−05	−12	−27	02	36	38	−40
Achievement Motivation	30	38	29	−38	33	12	12	20	33	−41	−24	13
Noncollege sample (N = 148)												
Optimism—Pessimism	13	18			13	18	23	02	02	−17	11	
General Adjustment	17	33			−18	30	38	02	15	−25	05	
Self-Confidence	29	36			30	28	38	18	12	−29	−08	
Affiliation	−15	−13			−17	03	00	−20	−02	10	41	
Work or Career Orientation	22	31			22	19	22	15	56	−11	02	
Leadership Role	38	51			36	35	46	21	28	−40	−32	
Dependence	−30	−42			−30	−28	−42	−19	−16	27	37	
Subordinate Role	−29	−40			−32	−24	−34	−20	−10	34	37	
Achievement Motivation	40	49			41	29	28	25	38	−50	−25	

with general effectiveness generally tend to be higher than are those with the staff predictions the patterns of relationship are quite similar.

The similarity of the matrices for the two samples, college graduate and noncollege, also is worth noting. The comparability of the correlations in the "Staff Prediction" column is particularly marked. The similarity was achieved despite differences in the educational backgrounds and occupational histories of the two samples and in spite of considerable turnover in the personnel of the assessment staffs. The finding implies that although the assessment process is judgmental and subjective considerable stability in the evaluations was achieved.

A clearer notion of the relative influence of the projective reports on the judgments of the assessment staffs can be obtained from Table 4. In this table the highest correlation with the staff evaluations of each assessment technique used in the program, other than the interview, is shown (the correlations for all techniques other than the projective reports were taken from Bray & Grant, 1966; the projective report correlations come from Table 3).

The correlations in Table 4 were inspected and the highest correlation with the scores based on each factor and the staff prediction identified. The latter correlations are italicized in the table.

It is clear from these data that the projective reports made a major contribution to the evaluation of motivational characteristics. For each of these characteristics, and for control of feelings in the noncollege sample, a projective variable has the highest correlation. It seems reasonable to infer from these data that the assessment staff was influenced by the projective reports, though not necessarily exclusively so, in their evaluations of the motivational characteristics of the men assessed.

That the influence of the projective reports on the overall evaluations of the assessment staffs was not as great as were those of other techniques also is evident from the data presented in Table 4. The highest correlations of any projective variable with general effectiveness (Factor I) and with the staff prediction are lower than those for most of the other techniques, particularly the Group Discussion and In-Basket. The projective correlations do compare favorably with those for the paper-and-pencil instruments, being about the same as those for the mental ability tests and higher than those for the personality questionnaires.

The correlations of the projective-report variables with other assessment techniques are generally low. A majority of the correlations fall below .20. Relatively few exceed .30 and only one (Achievement Motivation with

TABLE 4

HIGHEST CORRELATION OF EACH ASSESSMENT METHOD WITH STAFF JUDGMENTS

Assessment method	Staff prediction	Factor I	IV	V	VI	VII	VIII	IX	X	XI
College sample (N = 207)										
Manufacturing Problem	41	44	31	39	37	18	30	−35	−18	19
Group Discussion	*60*	*67*	48	*62*	*47*	31	45	−39	−22	31
In-Basket	55	60	*76*	45	39	36	44	−18	−15	17
Mental Ability Test	36	47	37	22	23	*79*	17	−16	−24	53
Personality Questionnaire	29	33	30	27	28	−23	28	−43	−29	−41
Projective Report	35	48	36	21	29	−33	*50*	*−47*	*49*	*−54*
Noncollege sample (N = 148)										
Manufacturing Problem	42	*60*	51	*52*	35	22	39	−43	−17	
Group Discussion	47	57	44	47	36	17	36	−41	−33	
In-Basket	*51*	59	68	49	24	27	26	−27	00	
Mental Ability Test	44	51	*72*	28	32	*64*	22	−22	−20	
Personality Questionnaire	24	22	30	−20	25	−19	18	−29	−34	
Projective Report	40	51	41	35	*46*	25	*56*	*−50*	*41*	

Note.—Numbers that are italicized indicate the highest correlation with the scores based on each factor and the staff prediction.

the Group Discussion in the noncollege sample) exceeds .40. In general, the correlations with other techniques are somewhat higher for the noncollege than for the college-graduate sample.

Correlations with other personality measures are of particular interest. In general, the nine projective variables have low correlations with the 15 "needs" measured by the Personal Preference Schedule (Edwards, 1953). Only 7 of the 135 intercorrelations exceed .20 in the college sample ($N = 207$) and 13 in the noncollege sample ($N = 148$). The highest correlations in each sample are .27, need Succorance with Dependence, college sample, and .36, need Nurturance with Subordinate Role, noncollege sample.

Correlations with the GAMIN (Guilford & Martin, 1943), obtained for the college sample only, tend to be somewhat higher than with the Personal Preference Schedule. Of the 45 intercorrelations, 10 exceed .20, all but 2 being in the .20s. The highest correlation is that of .37 between General Adjustment and Absence of Inferiority Feelings.

That the data from the projective reports are to some extent indicative of progress in management can be inferred from the correlations shown in Table 5. They represent average correlations for two groups of college-graduate managers with 7–9 years of service since being assessed and for two groups of noncollege managers with 8–9 years of service subsequent to assessment. The criterion, salary progress, was determined by computing the difference between salary at the time of assessment and salary on June 30, 1966. It is indicative of the progress in management made by these men up to the designated date.

It will be noted that the projective-report variables which correlate highest with the assessment-staff evaluations (Table 3) also tend to have the highest correlations with progress in management. This finding suggests

TABLE 5

CORRELATIONS WITH SALARY PROGRESS

Projective variable	College graduates (N = 81)	Noncollege (N = 120)
Optimism—Pessimism	−.01	.17
General Adjustment	.10	.19*
Self-Confidence	.11	.21*
Affiliation	−.06	−.07
Work or Career Orientation	.16	.17
Leadership Role	.24*	.19*
Dependence	−.35*	−.20*
Subordinate Role	−.25*	−.23*
Achievement Motivation	.26*	.30*

* $p < .05$ that $r = .00$.

that in favoring "achievement" and "leadership" motives to those of "adjustment" in evaluating managerial potential the assessment staffs were interpreting organizational values correctly.

Discussion

The above results indicate that the projective reports did make a significant contribution to the assessment of the participants in the Management Progress Study. The three projective variables which generally showed the highest correlations—leadership role, achievement motivation, and dependence—do not correlate as highly as do the situational techniques with the overall staff prediction, but they show about the same relationship to this prediction as do the mental ability measures. Their correlations are higher than are those of personality questionnaires.

The projective reports appear, as might be expected, to have particularly influenced the assessment staff in rating such characteristics as dependency, passivity, work motivation, and nonconformity. The relevant correlations are generally clearly higher than those between the personality inventories and the staff judgments.

Several of the projective variables are reliably related to salary progress. Those significant for both the college and noncollege groups are dependence, achievement motivation, leadership role, and subordinate role—the variables also most highly related to the overall staff prediction. There would seem to be little doubt that the projective reports do yield valid information regarding managerial motivation.

The conditions under which these encouraging results were obtained should be noted. The projective report is a summary report based on three different instruments. This report is in no way a scoring of the projective protocols but is impressionistic in nature. Neither is the report deeply "clinical"; every effort was made to orient it to the motivations relevant to business management. Finally, because four-fifths of the reports were written by one psychologist a question could be raised regarding the replicability of the findings.

REFERENCES

Bray, D. W. The management progress study. *American Psychologist*, 1964, **19**, 419–420.

Bray, D. W., & Grant, D. L. The assessment center in the measurement of potential for business management. *Psychological Monographs*, 1966, **80**(17, Whole No. 625).

Edwards, A. L. *Edwards Personal Preference Schedule*. New York: Psychological Corporation, 1953.

Guilford, J. P., & Martin, H. G. *The Guilford-Martin Inventory of Factors GAMIN*. Beverly Hills, Calif.: Sheridan Supply Co., 1943.

Hague, J. P., Otis, J. L., & Prien, E. P. Assessments of higher-level personnel: VI. Validity of predictions based on projective techniques. *Personnel Psychology*, 1962, **15**, 335–344.

Huse, E. F. Assessments of higher-level personnel: IV. The validity of assessment techniques based on systematically varied information. *Personnel Psychology*, 1962, **15**, 195–205.

Kelly, E. L., & Fiske, D. W. *The prediction of performance in clinical psychology*. Ann Arbor: University of Michigan Press, 1951.

Kinslinger, H. S. Application of projective techniques in personnel psychology since 1940. *Psychological Bulletin*, 1966, **66**, 134–149.

MacKinnon, D. W. An assessment study of Air Force officers. Part V: Summary and applications. *Technical Report WADC-TR-59-91 (V), ASTIA Document No. AD 210 220*. Washington, D. C.: Office of Technical Services, United States Department of Commerce, 1958.

(Early publication received January 23, 1967)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 233–235

VALIDITY OF THE JOB-CONCEPT INTERVIEW IN AN INDUSTRIAL SETTING

GERALD V. BARRETT
Goodyear Aerospace Corporation

BYRON SVETLIK
Western Reserve University

AND ERICH P. PRIEN
University of Akron

Interviews by psychologists were used to predict employee attitudes and job performance. Correlations were computed between the psychologist's predictions and employee ratings of their job attitudes and supervisor ratings of employee job performance. The psychologists were most accurate in predicting employee attitudes toward advancement and general morale. They were least accurate in predicting employee attitudes toward supervision and rewards. The psychologists were unable to predict employee performance as rated by the supervisors.

Considerable evidence regarding the validity of the interview for various purposes such as counseling, prediction of job performance, and prediction of consumer behavior has accumulated over the years. However, the reviews of Wagner (1949) and Mayfield (1964) suggest at best a marginal relationship between interviewer predictions and interviewee behavior. Certainly one of the primary limitations is the marginal relationship between what the interviewee says and does in the interview situation and his performance or behavior in future situations. Only a few studies have been designed to explore the interview process, although the work of Webster (1964) regarding the validity of interviewer predictions and the factors limiting validity has shown promising results.

The purpose of this study was to explore the relationship between predicted and measured employee job attitudes and predicted and rated job performance using the job-concept interview. These interviews were conducted in the actual work situation.

Method and Procedure

Three sources of data were available for this study: psychologists conducted job-concept interviews for the purpose of obtaining a job description and whatever other information and insight that could be obtained relevant to employee attitudes and behaviors; employees (predominantly clerks and secretaries) were requested to complete an attitude questionnaire consisting of six graphic rating scales; and supervisors were requested to complete a performance evaluation of each of their subordinates. The questionnaires completed by the interviewee and the interviewee's supervisor were identical to the questionnaire completed by the psychologist at the conclusion of the job-concept interview.

In sequence, the psychologist completed the job-concept interview making extensive notes during the process and immediately thereafter completed the questionnaire reflecting his appraisal of the interviewee's job attitudes and job performance. At a later time employees were asked to rate their job attitudes and at the same time the supervisors completed the performance evaluation form. The performance ratings from the supervisors' and the employees' rating of the attitude questionnaire were obtained with a psychologist present to answer any questions. Each employee identified himself voluntarily so that the analysis could be completed. Employees were informed that the data would be used exclusively for research purposes and that identities would never be disclosed. In that the psychologists had established relatively good rapport with the employees during the job-concept interviews, complete data were collected. Attitude questionnaires were not administered to those employees who were not at their regular place of work at the time of administration. There were no instances suggesting a lack of cooperation on the part of either employees or supervisors. The two questionnaires, one to measure the employee job attitudes and the other for the supervisor's ratings of subordinate's performance, were developed specifically for this research project and are described in a previous article by Svetlik, Prien, and Barrett (1964).

Product-moment correlations were computed between the psychologist's predictions of employee job attitudes and job performance and the employee responses to the attitude questionnaire, and the supervisor's job-performance ratings.

Results and Discussion

An examination of the results presented in Table 1 shows a positive and significant re-

TABLE 1

CORRELATION BETWEEN EMPLOYEE-ATTITUDE DIMENSIONS AND PSYCHOLOGIST'S PREDICTION OF ATTITUDES

Attitude dimensions	Psychologist's prediction of employee attitude							
	Supervision	Job	Management	Rewards	Advancement	Morale	*M*	σ
Employee attitude toward supervision	.06	.03	−.01	−.01	−.07	.02	3.88	.81
Job satisfaction	−.01	.23*	.08	.07	.24*	.25*	3.93	.95
Management and communications	−.04	.11	.22*	.12	.20	.07	3.53	.93
Rewards	.22*	.06	.12	.14	.17	.22*	2.83	1.11
Advancement	−.05	.02	−.08	.05	.30**	.10	3.11	1.06
General morale	.09	.20	.14	.03	.13	.29**	3.84	.92
M	3.35	3.49	3.13	2.99	3.12	3.17		
σ	.77	.83	.68	.60	.66	.68		

Note.—$N = 88$.
* $p < .05$.
** $p < .01$.

lationship between psychologist's predictions of employee job attitudes and the employee response to the attitude-questionnaire scales on four of the six scales. The degree of relationship varies and is most pronounced for attitude toward advancement and general morale and lowest for attitude toward supervision.

In predicting performance, the only appropriate correlation, as shown in Table 2, which approaches significance is that between the psychologist's prediction of job energy and the supervisor's rating of the employee on job energy. It should be noted that none of the psychologist's predictions of the three scales correlates significantly with the supervisor's rating of overall job performance. One inappropriate but significant correlation appears in the off-diagonal cells between the psychologist's prediction of job energy and the supervisor's rating of the employee job competence.

It would seem from the foregoing that, in the industrial situations, the face-to-face interview can serve a diagnostic purpose in predicting employee attitudes toward the job. The most valid prediction is for the dimensions of attitude toward advancement and general morale, and least valid for attitudes toward supervision and rewards. This result seems to be incongruous in that the interview focused on supervision received and super-

TABLE 2

CORRELATION BETWEEN SUPERVISOR RATINGS OF EMPLOYEE JOB PERFORMANCE AND PSYCHOLOGIST RATING OF PERFORMANCE

Supervisor ratings	Psychologist ratings				
	Personnel relations	Job competence	Job energy	$\bar{X}$	σ
Personnel relations	−.03	.17	.10	3.69	.87
Job competence	.06	.15	.24*	3.79	.80
Job energy	.00	.09	.19	4.08	.81
Overall effort	.11	.05	.01	3.57	.71
$\bar{X}$	3.31	3.58	3.62		
σ	.82	.84	.65		

Note.—$N = 95$.
* $p < .05$.

vision given within the job situation. The interviews did not contain specific questions to determine general morale, although advancement in both absolute and in relative terms was a factor of some considerable concern to the interviewee (this was a job-evaluation program). Many employees hoped that their job would be rated higher and that their pay would be increased as a direct result of what happened in the interview.

Somewhat disappointing and in contrast with some previous research by Otis and Daniels (1950), use of the job-content interview by a person other than the supervisor to appraise performance would not seem to be warranted.

Continued use of the interview as an information-gathering procedure seems to be assured. Use of the interview as a diagnostic procedure to predict the more immediate attitudes and perception of individuals is supported by this study. This study does not support the use of the interview as a technique for field-review purposes. In those situations where the field review is used, very likely the reviewer will require information obtained from sources other than the interview situation.

REFERENCES

MAYFIELD, E. C. The selection interview: A re-evaluation of published research. *Personnel Psychology*, 1964, **17**, 239–260.

OTIS, J. L., & DANIELS, H. W. A method for analyzing employment interviews. *Personnel Psychology*, 1950, **3**, 425–444.

SVETLIK, B., PRIEN, E. P., & BARRETT, G. Relationships between job difficulty, employee's attitude toward his job, and supervisory ratings of the employee effectiveness. *Journal of Applied Psychology*, 1964, **48**, 320–324.

WAGNER, R. The employment interview: A critical summary. *Personnel Psychology*, 1949, **2**, 17–46.

WEBSTER, E. L. Decision making in the employment interview. Industrial Relations Center, McGill University, 1964.

(Received June 9, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 236-241

A FACTORIAL STUDY OF SOME PSYCHOLOGICAL, VOCATIONAL INTEREST, AND MENTAL ABILITY VARIABLES AS PREDICTORS OF SUCCESS IN DENTAL SCHOOL

MARTIN K. CHEN, DALE W. PODSHADLEY
Division of Dental Health, United States Public Health Service, San Francisco
AND JOHN G. SHROCK[1]
University of California Medical Center, San Francisco

A total of 32 mental ability, past-achievement, manual skill, personality, and vocational interest variables believed to be potentially useful in the selection of dental students were factor analyzed to determine their factor pattern in relation to the criterion variable, the dental GPA. Then those variables which shared common factors with the criterion variable were used as independent variables in a multiple-regression equation for predictive purposes. For the 72 dental juniors studied, it was found that there were 2 common factors between the "predictor" variables and the dental GPA. 1 factor, decided to be academic aptitude, was significantly loaded in 4 "predictor" variables besides the dental GPA. The other factor, related to manual skill, was significantly loaded in 3 subtests of a manual skill test and the dental GPA. All but 1 of the 7 "predictor" variables were used in the predictive equation, 1 variable being dropped because of its low, though statistically significant, factor loading. Of all the variables, the predental GPA was found to be the most important predictor of success in dental school.

Since the Dental Aptitude Testing program was instituted by the American Dental Association some 20 years ago, numerous studies have been conducted to validate the test battery as predictors of achievement in dental school. Some of the studies, such as those by Weiss (1952) and Webb (1956), found moderate relationships between dental aptitude test scores and grades in certain courses. Other studies, including one by Layton (1953), found practically no relationship between the aptitude scores and grades in dental courses.

With respect to personality and vocational interest variables, Thompson (1944) found some low, though statistically significant, relationships between these variables and grades in the theory and technique courses in dental school. Graves (1942), on the other hand, found fairly high relationships between a psychological variable and dental student performance, but the relationships were far from stable from year to year. In view of the inconsistent and sometimes conflicting evidence of the validity of these variables as predictors of achievement in dental school, the present study was undertaken to explore a large battery of potentially useful variables by factor analysis prior to the application of the multiple-regression technique for prediction. This procedure accords with the suggested uses of factor analysis mentioned by Harman (1960, p. 7).

Purpose of Study

The purpose of the study is twofold: (*a*) to explore, by factor analysis, a large number of psychological, vocational interest, and prescholastic ability variables as potential predictors of success in dental school; and (*b*) to derive a multiple-regression equation, using as independent variables those variables which shared common factors with the criterion of success, the dental grade-point average (GPA). According to Guilford (1954, p. 356), for a predictor variable to have any

[1] The authors gratefully acknowledge the assistance of Henry F. Kaiser, who reviewed the final draft of this paper and made several valuable suggestions which have been incorporated into the paper.

validity at all, it must have at least one common factor with the criterion variable. Thus the criterion of selection of the independent variables for the multiple-regression equation was that they have at least one common factor with the dependent or predicted variable.

PROCEDURE

The entire sophomore class of 72 students enrolled in the University of California School of Dentistry in the academic year 1964–1965 was used as subjects (*Ss*) in the study. The *Ss*' Dental Aptitude Test (DAT) scores, their predental grade-point averages (predental GPA), and their scores on the California Performance Test (CPT)[2] were obtained from the students' records. The DAT scores and CPT scores were routinely required of all applicants to the school of dentistry.

In the summer of 1964, the California Psychological Inventory (CPI) and the Kuder Preference Record, Form D, were administered to this class. Finally, *Ss*' GPAs cumulative to the end of the first semester of the junior year (dental GPA) were obtained from their records. Data analysis was made by two computer programs,[3] one computing the intercorrelations and the other performing the factor analysis by the principal component method, followed by Quartimax and Varimax rotations.

Thus, for this study, the correlation matrix of 33 variables included the 18 scales of the CPI; the predental GPA; the nine subtests of the DAT on quantitative reasoning, verbal reasoning, reading comprehension in the natural sciences, biology, chemistry, factual science information, science application, and carving ability; the dentist scale of the Kuder Preference Record; the three subtests of tooth, plaster, and wax of the California Performance Test; and the dental GPA, the last being the criterion variable. Data were available from 68 of the 72 *Ss*, though for several *Ss* the data were incomplete. Since the *R* technique (correlation between traits) rather than the *Q* technique (correlation between persons) was used, the unequal *N*s presented no particular problem for the computer programs which were written to accommodate missing data.

FINDINGS AND INTERPRETATIONS

Since the mathematical statisticians have so far failed to come up with an exact standard-error formula for testing the reliability of factor coefficients, the approximation formula developed by Holzinger and Harman (1941, p. 417) was used. According to this formula, the standard error in the sampling distribution of factor coefficients for $r = .18$ and $N = 68$ is approximately .214, where rho is estimated by the average of the correlation coefficients included in the matrix for factoring. Thus, for a factor to be significant at the .05 level, it must be at least $2 \times .214 = .428$.

Of the 496 correlation coefficients that entered into the matrix, 90 or 18.1% were significant at the .01 level, 45 or 9% significant at the .05 level, and 30 or 6% significant at the .10 level. This means that 165 or 33.2% of all the coefficients possess some degree of significance, giving a measure of assurance that the factors have not been inordinately attenuated by the unreliability of the correlation coefficients.

Altogether 12 principal components (factors) were extracted. Rotations by the Varimax resulted in 9 interpretable orthogonal factors, whereas by the Quartimax method 11 interpretable orthogonal factors emerged. Both Kaiser (1958) and Wrigley, Saunders, and Neuhaus (1958) have found that the Varimax solution results in factors that have greater invariance and that are closer to simple structure. For this reason, interpretations of the factors in this study are based on the Varimax solution.

Of the nine factors, only two had significant loadings in the dental GPA, the criterion variable. The coefficients of the two factors in the 33 variables are given in Table 1. Factor I is heavily loaded in the three subtests of the California Performance Test, but the loading is barely significant in the criterion variable.

The tooth, plaster, and wax subtests of the CPT are tests of manual skills in shaping a piece of clay into an oversized tooth, carving a cube of plaster in a specified manner, and

[2] The California Performance Test consists of three subtests on tooth molding, chalk carving, and filling outline forms with melting wax. Developed by the Admissions Committee of the University of California School of Dentistry, this test has been used together with the carving-ability test of the Dental Aptitude Test for over 10 years. No information on the test's reliability and validity is available.

[3] Both programs were developed by the Institute of Human Development, University of California, Berkeley. The MCO does *Q* or *R* correlations and tests the coefficients at the .01, .05, and .10 levels. The FARO uses the principal component method in factoring the correlation matrix prepared by MCO and performs orthogonal rotations by both the Quartimax and Varimax solutions.

TABLE 1

MATRIX OF TWO FACTORS WITH SIGNIFICANT LOADINGS IN THE CRITERION VARIABLE (DENTAL GPA)

	Factor I	Factor II
Percentage of Communality	9.940	10.500
Dental GPA	.430*	.694*
Predental GPA	−.122	.803*
Kuder Preference Record, Form D	.148	−.024
California Psychological Inventory		
Dominance	−.016	−.011
Capacity for Status	−.012	.070
Sociability	.023	−.105
Social Presence	−.086	−.108
Self-Acceptance	−.018	.051
Well-Being	.036	.030
Responsibility	.032	.222
Socialization	−.073	.040
Self-Control	.009	.042
Tolerance	.137	.135
Good Impression	−.051	−.128
Communality	.121	.118
Achievement via Conformance	−.056	.431*
Achievement via Independence	.054	.221
Intellectual Efficiency	−.061	.350
Psychological-Mindedness	.005	.231
Flexibility	.038	−.085
Femininity	.023	.013
Dental Aptitude Test		
Quantitative Reasoning	−.348	.443*
Verbal Reasoning	−.121	.625*
Reading Comprehension	.049	.190
Biology	.371	−.145
Chemistry	.019	.212
Factual Information	.143	.081
Science Application	.144	.045
Space Relations	−.084	−.029
Carving Ability	.301	.075
California Performance Test		
Tooth	.688*	−.062
Plaster	.732*	.087
Wax	.749*	−.077

* Significant factor coefficients.

melting a stick of wax into simple outlines on a plaster slab. The high coefficients of this factor in the tests indicate that Factor I is a manual skill factor, which is obviously required in the technique courses of dentistry.

It is interesting to note that the carving-ability test of the DAT, which purports to measure manual dexterity, has a nonsignificant loading of the manual skill factor (.301). This would seem to suggest that this test is not a measure of digital dexterity. An examination of the complete nine-factor matrix shows that the ninth factor is heavily loaded in the carving-ability test (.713) and moderately loaded in the Cm (communality) scale of the CPI (−.633). The next lower coefficient of this factor is found in the Py (psychological-mindedness) scale of the CPI (.410), which is not statistically significant. The significant positive and negative loadings of this factor in the carving-ability test and the Cm scale indicate that it is a bipolar characteristic. What this characteristic is can only be a matter of speculation, since neither the carving-ability test nor the Cm scale yields much information in this respect. The descriptive terms of the Cm scale, however, such as steady, patient, and moderate for high scorers, and impatient, disorderly, and nervous for low scorers, suggest that this is a dispositional or temperamental factor. Credence is lent to this interpretation by dentists who feel the typical chalk-carving testing situation generates so much tension that the scores on the test could very well reflect the testees' affective responses to stress rather than their true digital abilities.

Table 1 shows that Factor II has a high loading in the predental GPA, moderate loadings in verbal reasoning of the DAT and the dental GPA, and low loadings in quantitative reasoning of the DAT and the Ac (achievement via conformance) scale of the CPI. The verbal and quantitative reasoning tests of the DAT are based on the V (verbal) and Q (quantitative) scores of the Cooperative School and College Ability Tests (SCAT). The SCAT manual (Educational Testing Service, 1957, p. 6) claims that the V and Q subscores are measures of "developed abilities" that are related to success in school learning. It thus appears that Factor II is the academic-aptitude factor. This interpretation is reinforced by the factor's significant loading in the Ac scale, which is said to measure traits related to intellectual achievement in a milieu where conformance is a requisite.

Except for the verbal and quantitative reasoning tests, the Dental Aptitude Test, which all applicants to a dental school are required to take, has no validity as a predictor of success for the sample studied, insofar as it has no common factors with the dental

GPA. It appears that the entire battery, with the exception of the verbal and quantitative reasoning tests and the carving test, measures one important factor, which is loaded neither in the predental GPA nor in the dental GPA. The coefficients of this factor, which are given in Table 2, show that so far as this one factor is concerned, one of the subtests would do practically just as well as the entire battery.

An examination of the reading-comprehension, biology, chemistry, factual information, and science application tests in which this factor is loaded, reveals that recall would be a major factor for success in these tests. On the other hand, recall plays no role in the space-relations test, in which this factor also has a sizable loading. Thus the common factor could not be recall. Further reflection on the nature of the tests brings to light the fact that the speed with which relationships between parts are recognized would be required to do well in both the space-relations and the other tests, all of which are timed. For this reason this factor is tentatively termed cognitive facility.

No attempt will be made to interpret the other factors in the nine-factor matrix, since they are not relevant to this study. It will be pointed out, however, that for this sample studied, the CPI and the dentist scale of the Kuder Preference Record are practically useless as predictors of student performance in dental school. As a matter of fact, out of the 32 predictor variables used in this study, only the three subtests of the CPT, the predental GPA, the quantitative and verbal reasoning of the DAT, and the Ac scale of the CPI have any predictive value so far as achievement in dental school is concerned.

One way of looking at the efficiency of the predictor variables is to study the valid predictor variables in relation to the two pertinent factors. For Factor I, the manual dexterity variable, the pertinent variables are the three subtests of the CPT and the dental GPA. For the tooth subtest, the square of the factor coefficient is .473, that for the plaster subtest is .536, that for the wax subtest is .561, and that for the criterion variable is .185. This means that while this factor accounted for the variances of the three CPT subtests to the extent ranging from 47.3% to 56.1%, it only accounted for the variance of the criterion variable to the extent of 18.5%. In other words, 81.5% of the variance of the criterion variable was *not* accounted for by this one factor.

For Factor II, the academic-aptitude factor, the relevant variables are the predental GPA, verbal and quantitative reasoning of the DAT, the Ac scale of the CPI, and the dental GPA. The squares of their factor coefficients are: .645, .391, .196, .186, and .482, respectively. In this case, the factor accounted for 48.2% of the variance of the criterion vari-

TABLE 2

COEFFICIENTS OF THE COGNITIVE FACILITY FACTOR COMMON TO TESTS OF THE DAT BATTERY

	Factor
Percentage of Communality	17.372
Dental GPA	.084
Predental GPA	.114
Kuder Preference Record, Form D	.056
California Psychological Inventory	
Dominance	−.135
Capacity for Status	.079
Sociability	.006
Social Presence	.128
Self-Acceptance	−.025
Well-Being	−.078
Responsibility	−.172
Socialization	−.109
Self-Control	.153
Tolerance	−.133
Good Impression	.257
Communality	−.160
Achievement via Conformance	−.148
Achievement via Independence	.144
Intellectual Efficiency	−.117
Psychological-Mindedness	.322
Flexibility	.355
Femininity	−.122
Dental Aptitude Test	
Quantitative Reasoning	.295
Verbal Reasoning	.386
Reading Comprehension	.666*
Biology	.629*
Chemistry	.836*
Factual Information	.671*
Science Application	.852*
Space Relations	.587*
Carving Ability	.198
California Performance Test	
Tooth	.170
Plaster	.237
Wax	−.110

* Significant factor coefficients.

able, and 64.5% of the variance of the predental GPA, indicating that, of all the predictor variables, the predental GPA is the most efficient.

Confirmation of this statement was given by the multiple-regression equation, in which all but one of the seven variables which shared common factors with the dental GPA were used as independent variables. The one variable, the Ac scale of the CPI, was arbitrarily dropped because of its low, though statistically significant, factor loading. In the equation

$$X' = .042X_2 + .084X_3 + .042X_4 + .480X_5 - .016X_6 + .036X_7 + .540,$$

X' is the predicted dental GPA; X_2 the tooth subtest, X_3 the plaster subtest, and X_4 the wax subtest of the CPT; X_5 the predental GPA; X_6 quantitative reasoning, and X_7 verbal reasoning of the DAT. The relative importance of the independent variables in terms of their contributions to the prediction can be judged by the squares of the beta coefficients, which, in the order in which the variables entered into the equation, are: .018, .067, .024, .305, .008, .034. It is evident that X_5, the predental GPA, is by far the most important of all the predictors for this study.

It may be of interest to find out how much predictive power was lost by using 6 of the battery of 32 potential predictor variables rather than the entire battery. Dwyer (1939) has shown that an approximate estimate of this loss may be obtained by comparing the multiple correlation obtained with the square root of the communality of the predicted variable in the factor matrix. The multiple correlation between the predicted variable and the 6 independent variables is .719, and the square root of the communality of the dental GPA is .860, which would be the maximum multiple correlation between the dental GPA and all the 32 variables if they were used. Thus the loss of predictive power is in the order of .141, in terms of the multiple correlation.

Discussion

A total of 33 variables was included in the correlation matrix from which 12 principal components were extracted. Orthogonal rotations for an analytic solution by the Varimax method resulted in 9 interpretable factors. Of the 9 factors, only 2 were found to have significant loadings in the criterion variable, the dental GPA. Factor I is identified as a manual skill factor, with substantial loadings in the three subtests of the CPT. Factor II, termed the academic-aptitude factor, has significant loadings in the predental GPA, the quantitative and verbal reasoning tests of the DAT, the Ac (achievement via conformance) scale of the CPI, as well as the dental GPA. It was further decided that the carving-ability test of the DAT measured a factor related to temperament rather than simple digital skill, and that the entire DAT battery, with the exception of quantitative and verbal reasoning and carving ability, measures one common factor suggested to be cognitive facility. The predental GPA was found to be the most efficient predictor of the predictor variables, both in terms of its common factor loading with the dental GPA and in terms of its contribution to the multiple-regression equation.

It must be remembered that the findings apply only to the sample studied. Whether or not the results will remain invariant for other samples from the same institution cannot be predicted until results of further research are known. Travers and Wallace (1950), for example, have found that the DAT battery have little value as predictors for one class in the University of Michigan School of Dentistry, but considerable value for another class. This phenomenon necessitates replications of the same study with succeeding classes in one institution, so that the effects of class peculiarities can be evaluated and eliminated from consideration in the analysis of data.

It is necessary, too, to keep in mind that at best the GPA is an unsatisfactory criterion of academic success in any college or professional school. Many of the grades which enter into the computation of the GPA are not objectively derived, and as such they reflect, in large measure, the instructor's personal prejudices rather than the students' true achievement. Because of this fact the GPA is likely to be unstable, depending as it does on the judgments of the instructors con-

cerned. This instability will result in the attenuation of its correlations with other measures, and consequently in the attenuation of the coefficients of common factors with these measures. This fact may, in part, explain the invalidity of the predictor variables, particularly the DAT battery, in this study.

It has been shown that the academic-aptitude factor is more important to student achievement in dental school than the manual skill factor. The reason for this is perhaps that the first $2\frac{1}{2}$ years of the typical dental curriculum comprises more academic courses than technique courses. This automatically weights the two categories of courses in the computation of the cumulative dental GPA, and the difference in the magnitude of the coefficients of the two factors probably reflects the consequence of this differential weighting. A more definitive statement about this could have been made had it been possible to secure data on the dental GPA dichotomized according to the nature of the courses that entered into its computation.

REFERENCES

DWYER, P. S. The contribution of an orthogonal multiple factor solution to multiple correlation. *Psychometrika*, 1939, **4**, 163–171.

EDUCATIONAL TESTING SERVICE. *Cooperative School and College Ability Tests, manual for interpreting scores.* Princeton, N. J.: Author, 1957.

GRAVES, H. A. Factors in dental aptitude I. *Proceedings of American Association of Dental Schools,* 1942, **19**, 252–257.

GUILFORD, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

HARMAN, H. H. *Modern factor analysis.* Chicago: University of Chicago Press, 1960.

HOLZINGER, K. J., & HARMAN, H. H. *Factor analysis.* Chicago: University of Chicago Press, 1941.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika,* 1958, **23**, 187–200.

LAYTON, W. L. Predicting success in dental school. *Journal of Applied Psychology,* 1953, **37**, 251–255.

THOMPSON, C. E. Personality and interest factors in dental school success. *Educational & Psychological Measurement,* 1944, **4**, 299–306.

TRAVERS, R. M. W., & WALLACE, W. L. Inconsistency in the predictive value of a battery of tests. *Journal of Applied Psychology,* 1950, **34**, 237–239.

WEBB, S. C. The prediction of achievement for first-year dental students. *Educational & Psychological Measurement,* 1956, **16**, 543–548.

WEISS, I. Prediction of achievement success in dental school. *Journal of Applied Psychology,* 1952, **36**, 11–14.

WRIGLEY, C., SAUNDERS, D. R., & NEUHAUS, J. O. Application of the quartimax method of rotation to Thurstone's primary mental abilities study. *Psychometrika,* 1958, **23**, 151–170.

(Received June 13, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 242-246

SELF AND OTHER SEMANTIC CONCEPTS IN RELATION TO CHOICE OF A VOCATION [1]

RICHARD A. HUNT [2]

Texas Christian University

Criterion groups (engineers, managers, ministers, teachers) composed of 258 professional men judged 16 self-, other, and vocational concepts on 25 semantic differential scales. Concept interrelationships were measured with a modified, normalized D score. A multiple-discriminant analysis produced 3 significant functions (each $p < .001$) between the criterion groups. In a cross-validation group of male undergraduates, semantic differential profiles correctly classified 70 of 139 Ss according to Kuder Occupational Interest Inventory (OII) scores and 83 of 125 of these Ss according to 1st choice of vocation. Complex, real-life decisions can be predicted from individual patterns of self- and other concepts. Results support assumptions about the self-concept in the theories of Rogers and Super.

An adequate understanding of how an individual selects a particular pattern of behavior from among several alternatives is a fundamental problem. Erikson (1950, 1959) and Rogers (1959) have emphasized the influence of an individual's self-concept upon his complex decisions. Super (1957) has attached great significance to the role of the self-concept in the development of an individual's vocational choice. Cronbach (1960) states that vocational interests are an expression of personality needs and patterns.

The theoretical issue of this research was whether complex decisions are systematically related to self-concepts. The basic purpose was to investigate the relation between an individual's self-concepts and his vocational choice as one empirical manifestation of the relation between personality and the decision process.

The emphasis of the following hypotheses is upon the pattern of interrelationships between semantic concepts which an individual may hold about himself, specific persons in his life, and "contrived, stereotype" persons such as "the typical engineer."

Hypothesis 1. The relationship between the self-concept and other semantic concepts will be differentially related to vocational choice.

Hypothesis 2. The vocational interest of individuals can be predicted from the interrelationships between their self-concepts and other concepts.

Method and Procedure

Semantic Differential Instrument

Concept similarity was measured by using the semantic differential technique (Osgood, Suci, & Tannenbaum, 1957). However, as a conceptual aid in relating the semantic differential scores to Rogers' (1959) theory, Osgood's *D* score was changed to a normalized convergence (*C*) score. After the usual *D* scores between all pairs of concepts were secured for an individual case, the *C* score was calculated between Concepts i and j for an individual matrix by the following formula:

$$C_{ij} = \left(\frac{(D_{ij} - \bar{D})}{\sigma_D} \cdot (-2)\right) + 10$$

In this scoring system a higher score indicated greater similarity or congruence between a pair of concepts. In addition, the scores for an individual case were normalized within that case with a mean of 10 and a standard deviation of 2.

The semantic differential instrument which was used consisted of 16 concepts. These concepts were worded in a manner which would make them equally familiar and appropriate to all of the subjects (*Ss*) in the study. The exact wording of these

[1] This report is based upon the author's dissertation in partial fulfillment of requirements for the doctoral degree at Texas Christian University. The author expresses his appreciation to Winton H. Manning, dissertation chairman, and the faculty of the psychology department at Texas Christian University for helpful suggestions and encouragement in this research. Alvin J. North and Jack R. Strange, both of Southern Methodist University, have given valuable criticism of the present manuscript.

[2] Now at Southern Methodist University.

concepts as they appeared in the test booklet is provided below.

1. My acting self: How I intend to act or behave toward other people.
2. My outer self: How other people usually see me and think of me.
3. My inner self: How I really feel about myself deep within my innermost thoughts.
4. My ideal self: How I would like to be if I had the opportunity.
5. My father (or if you never knew your father, the man who was most like a father to you).
6. My mother (or if you never knew your mother, the woman who was most like a mother to you).
7. Most men: The average or typical man.
8. Most women: The average or typical woman.
9. God: My idea of the Supreme Being.
10. The average or typical minister: Men who work in a local church as a pastor.
11. The average or typical engineer: Men who work in industry in a mechanical, electrical, chemical, or other engineering speciality.
12. The average or typical teacher: Men in junior or senior high school who teach mathematics or science.
13. The average or typical manager: Men in business or industry who supervise other employees.
14. My chosen vocational field is ______: I personally think most people who are engaged in this field are:
15. My chosen vocational field is ______: I personally think most people who are *not* professionally trained in this field (not engaged in this vocation) are:
16. The person who has influenced me most in my life is ______. He (she) sees me as:

A total of 135 possible scales composed of adjectives or adjective phrases were initially constructed or drawn from previous research (Jenkins, Russell, & Suci, 1958; Osgood et al., 1957). Of these, a set of 90 was used with 89 students who were enrolled in introductory psychology classes. In this pilot study *Ss* judged only the vocational concepts. On the basis of these results the 25 scales which showed greatest variability and differentiation between the vocational concepts were selected for the experimental instrument.

The selected scales were: rushed-leisurely, democratic-autocratic, selfish-unselfish, idealistic-realistic, helpful-very helpful, other centered-self centered, very important-important, vague-precise, competitive-cooperative, very valuable-valuable, complacent-curious, independent-dependent, tender-tough, very good-good, average-outstanding, relaxed-tense, poorly paid-well paid, listening-talkative, generous-thrifty, proud-humble, contented-creative, thing centered-person centered, excitable-calm, compromising-independent, novel-routine.

One may question whether some of these scales are as truly bipolar as Osgood et al. (1957) require. However, the assumption concerning semantic scales which supports this selection procedure is that these scales are actually nominal representations of an underlying infinite continuum. For example, the supposed end points of the scale *good-bad* can always be extended to new end points by the use of superlatives. The scale *good-very good* actually represents a magnification of a segment of the *good-bad* scale. In the preliminary study *Ss* used all seven spaces of the adjective-superlative scales (e.g., *good-very good*), but with the adjective-opposite pairs *Ss* used only the three spaces nearer the favorable adjective (e.g., the "good" end of the *good-bad* scale). However, in every case the ordinal placement of the concepts as judged on the two related scales was the same.

In the final semantic differential the 25 scales were counterbalanced for left versus right positions of the adjectives and for order from top to bottom of the page. The concepts were arranged in random order within each of these clusters: (*a*) the four self-concepts (Concepts 1–4), (*b*) the five significant-person concepts (Concepts 5–9), (*c*) the four vocational concepts (Concepts 10–13), and (*d*) the three individualized concepts (Concepts 14–16).

TABLE 1

DESCRIPTION OF OCCUPATIONAL SAMPLES

Occupational group	Age		Years in occupation		Number who had achieved this educational level[a]					Total *N*
	M	*SD*	*M*	*SD*	1	2	3	4	5	
Engineer	31.23	5.67	6.51	4.71	0	10	47	5	3	65
Manager	40.20	9.00	12.75	8.38	14	14	22	14	5	69
Minister	40.17	11.49	13.88	10.95	7	5	12	35	7	66
Teacher	34.62	9.25	9.40	9.63	0	10	23	9	16	58

[a] Explanation of educational level:
1. Attended college but did not graduate from college.
2. Graduated from college with a baccalaureate degree.
3. Attended professional graduate school or graduate credit toward a master's degree.
4. Graduated from seminary or graduate school with the bachelor of divinity or master's degree.
5. Further graduate study beyond the master's or bachelor of divinity level.

TABLE 2

MEAN C SCORES FOR ACTING SELF-CONCEPT VERSUS VOCATIONAL CONCEPTS

Concept comparison	F ratio	Occupational group			
		Minister	Engineer	Teacher	Manager
Acting Self versus					
Minister	3.86*	10.33	9.58	10.06	9.48
Engineer	11.73**	8.87	10.39	9.36	9.11
Teacher	3.93**	9.91	10.04	10.66	9.80
Manager	8.15**	8.33	9.07	8.61	9.69

* $p < .02$
** $p < .01$.

Subjects

Four vocational criterion groups were composed of men who had completed basic formal academic preparation for their respective professions and were engaged in that vocation full time. Table 1 provides a description of these samples.

The self-explanatory semantic differential instrument was distributed at professional meetings and by mail. There were a total of 258 usable instruments returned by the four criterion groups.

A validation group consisted of 151 undergraduate students from freshmen through seniors who were enrolled full time at Texas Christian University. In addition to the semantic differential, these *Ss* also completed the Kuder (1957) Occupational Interest Inventory (OII), Form D. This was scored on the verification key and seven occupational keys: electrical engineer, mechanical engineer, high school mathematics teacher, high school science teacher, minister, personnel manager, and insurance agent. Since no data were available on these *Ss'* postcollege vocational decisions, scores on these scales were taken as the criterion of decision against which to validate the predictions from the semantic differential data.

RESULTS

Preliminary comparison of univariate analyses of variance across the criterion groups on each of the pairs of concept comparisons showed much better discrimination between the groups on the basis of the C scores than with the unmodified D scores. Out of 136 comparisons the F ratio for the C measure on a given comparison was greater than the F ratio for the corresponding D measure in 114 instances. There were 32 instances in which the F ratio for the C measure led to statistically significant group differences (p = .05 or less), but the D measure failed to indicate group differences at the .05 level. Since the C measure provided for finer discriminations between groups by eliminating variability due to *Ss'* means and *SDs*, it was used in the analysis of the data.

Hypothesis 1

The C scores for the acting self-concept versus the four vocational concepts in the four criterion groups are presented in Table 2. Within each of the four criterion groups there are significant differences between the C scores for the acting self-concept as compared with each of the four vocational concepts.

In each criterion group the highest C score of congruence or similarity occurs for the acting self-concept in comparison with the vocational concept which represents that group. Members of each vocational group tend to see themselves as more like members of their own group than like members of the other vocational groups in this study. For instance, in the minister group the highest C score mean is between acting self-concept versus minister concept.

Not only the differential relation of concept comparisons with groups is supported, but also the higher congruence score of the acting self-concept versus the specified vocational group concept presents a stable and predictable pattern. The same pattern of congruence scores appeared between the four vocational concepts as compared with the concepts of outer self, inner self, ideal self, influential person, and people in my vocational group. This pattern does not appear for the remaining concepts as compared with the four vocational concepts.

In a similar analysis each semantic self-

concept (Concepts 1–4) was compared within each criterion group with the concepts of father, mother, and persons within and outside of the individual's vocation (Concepts 5, 6, 14, 15). No differences appeared between the criterion groups in the C scores for each of these concept comparisons.

In each of the four criterion groups the C score for the self-concept versus the father concept is higher (the concepts are more similar) than the self-concept versus the mother concept. The probability of this pattern occurring by chance for the 16 comparisons (four self-concepts in each of four groups) is $\frac{1}{2}^{16}$ or .00003. The C score between the concepts of self versus the concept of typical persons in my vocation was also higher in each of these 16 instances of comparison than the same self-concepts versus the concept of persons who are not in my vocation.

In addition to the univariate tests of Hypothesis 1, a multiple-discriminant analysis (Cooley & Lohnes, 1962) of differences between the four criterion groups was performed. The C scores for 40 pairs of concepts selected from the 16×16 matrix of concept comparisons were used as the variables in this analysis. These variables were the four vocational concepts versus the concepts of inner self, ideal self, influential person, people in my vocational group, outer self, acting self, father, mother, God, and comparisons between the vocational concepts. The results of this analysis are shown in Table 3. This analysis indicates that the interrelationships between concepts for each of the criterion groups provides clear differentiation between these four occupational groups.

TABLE 3

Significance of the Discriminant Function Chi-Square Approximations

Function	Percentage of variance	df	χ^2
I	63.37	42	273.145*
II	22.49	40	135.320*
III	14.14	38	93.585*

Note.—Wilks Lambda = .119, $F = 5.559$, $df = 120$ 645, $p < .00001$.
* $p < .001$.

TABLE 4

Comparison of Semantic Differential and Kuder Occupational Interest Inventory Classifications in Validation Group

Semantic differential classification	Kuder OII, Form D			
	Minister	Engineer	Teacher	Manager
Minister	18.0	.5	2.5	6.0
Engineer	4.5	23.0	7.5	6.0
Teacher	9.0	5.0	7.5	7.5
Manager	3.0	11.0	6.5	21.5

Note.—$\chi^2 = 57.99$, $p < .001$, $N = 139$, Hits = 70.

Hypothesis 2

On the basis of the multiple-discriminant analysis of the criterion groups, a classification analysis (Cooley & Lohnes, 1962) of the 151 undergraduate students in the validation group was made. Table 4 compares the semantic differential classifications with the Kuder OII probabilities of group membership (Kuder's differential ratio, DR, score). In this classification procedure each S was placed in the occupational group for which his probability of group membership was highest. The seven scales of the Kuder were reduced to four by allowing the electrical and mechanical engineering scales to represent the engineering group, the teacher of mathematics and teacher of science in high school to represent the teacher group, and the personnel manager and insurance agent to represent the manager group. The Kuder minister scale represented the minister group. There were ties between the Kuder DR scores for 17 cases. These were broken by assigning one-half case to each of the two occupations involved.

The 70 hits out of 139 usable cases indicate that statistically dependable prediction of interest and probable choice of a vocation is possible on the basis of the interrelationships between the semantic differential concepts. The contingency coefficient of correlation between the discriminant-analysis classification and the Kuder classification is .543 compared to C_{max} of .866.

The discriminant-analysis classification of Ss in the validation group was also compared

TABLE 5

COMPARISON OF SEMANTIC DIFFERENTIAL CLASSIFICATION AND FIRST CHOICE OF VOCATION IN VALIDATION GROUP

Semantic differential classification	First choice of vocation			
	Minister	Engineer	Teacher	Manager
Minister	22	0	0	2
Engineer	0	26	7	7
Teacher	1	6	13	4
Manager	3	9	3	22

Note.—$\chi^2 = 139.86$, $p < .001$, $N = 125$, Hits $= 83$.

with their first choice of occupation. Of 125 *S*s whose first choice of vocation was one of the four in this study, 83 were correspondingly classified on the basis of their semantic differential *C* scores. The contingency coefficient of correlation between these two variables is .728 with $C_{max} = .866$.

DISCUSSION

Greater similarity or congruence between an individual's perception of himself and his perception of specific alternatives open to him (such as vocational types or areas) is related to his choice of a given alternative or vocation. This tends to support Rogers' (1959) theory that an individual attempts to maximize the congruence between awareness, experience, and behavior in making his decisions. As suggested by Super (1957), an individual tends to express his self-concept through his complex, real-life decisions, such as in the case of vocational choice and decision.

The present data provide clear evidence of the importance of the self-concept in relation to decisions in the natural environment. Future research must explore more exactly the psychological nature of these relationships in both theoretical and applied terms.

REFERENCES

COOLEY, W. W., & LOHNES, P. R. *Multivariate procedures for the behavioral sciences.* New York: Wiley, 1962.

CRONBACH, L. J. *Essentials of psychological testing.* (2nd ed.) New York: Harper, 1960.

ERIKSON, E. H. *Childhood and society.* New York: W. W. Norton, 1950.

ERIKSON, E. H. Identity and the life cycle. *Psychological Issues*, 1959, 1, 1–171.

JENKINS, J. J., RUSSELL, W. A., & SUCI, G. J. An atlas of semantic profiles for 360 words. *American Journal of Psychology*, 1958, 71, 688–699.

KUDER, G. F. *Research handbook for the Kuder Occupational Preference Record, Form D.* Chicago: Science Research Associates, 1957.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. *The measurement of meaning.* Urbana, Ill.: University of Illinois Press, 1957.

ROGERS, C. R. A theory of therapy, personality, and interpersonal relationships as developed in the client-centered framework. In S. Koch (Ed.), *Psychology: A study of a science.* Vol. 3. *Formulations of the person and the social context.* New York: McGraw-Hill, 1959.

SUPER, D. E. *Vocational development: A framework for research.* New York: Columbia University Press, 1957.

(Received June 13, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 247-253

FACTORS IN COLLEGE ATTENDANCE[1]

ARTHUR A. DOLE AND JOHN M. DIGMAN
University of Hawaii

Purpose of this research was to identify factors, obtained from a self-report inventory, determining the decision to attend college. 214 male and 306 female students from the state university completed a 78-item inventory, Reasons for Going to College, as freshmen and again as seniors. Analyses and matching procedures yielded 13 major factors: Social Reason, Conformity, Curiosity, Vocational Reason, Academic Value, Material Value, Altruistic Value, School Influence, Experience, Avocational Influence, Science Interest, Humanities Interest, and Verbal Interest. Additional factor analyses of a dismissed college group and 3 diverse public-school samples suggested that these major factors had considerable generality. Implications for educational decision making and for further research were considered.

It is well known that the decision to enter college is associated with occupational goal, socioeconomic status, family aspirations, birth order, proximity of college, secondary school attended, educational achievement, ability, and other variables (Berdie, 1954; Flanagan, Davis, Dailey, Shaycoft, Orr, Goldberg, & Neyman, 1964; Iffert, 1957; McClelland, Baldwin, Bronfenbrenner, & Strodtbeck, 1958; Sanford, 1962). Among investigations of motivation for higher education, those of Iffert (1957) and Goldsen, Rosenberg, Williams, and Suchman (1960) are most pertinent to the present study. A decade ago Iffert attempted, unsuccessfully, to relate self-reported "reasons for going to college" to withdrawal or continuation in higher education. Goldsen et al. have shown that values (e.g., materialism and other-directedness) are associated with curricular preferences. In addition, the work of Spindler (1955), Bidwell, King, Finnie, and Scarr (1963), and Getzels (1957) can be cited in support of a traditional-versus-emergent value dichotomy. This research was concerned with the identification of major factors in reasons given for the decision to attend college. It sought to show the extent to which such factors are stable and generalizable to various educational groups.

Problem

After a review of the literature, interviews with students, and a series of open-ended and structured (checklist) surveys of undergraduate groups, Dole (1965b) proposed that the reported determinants for entering college might sensibly be classified as Reasons (Iffert, 1957), Values, Interests, and Influences. Dole also concluded that the motivational factors underlying the decision to go to college were probably quite complex. An investigation of the determinants of choice at six educational levels, involving a total of more than 20,000 subjects (*Ss*) within a single state (Dole, 1964a, 1964b), suggested further that the basic classification scheme might be extended to a wide variety of decision situations.

In the present study the problem was approached through factor analysis. Specifically, it was hoped to (*a*) identify major factors in reasons given for college attendance; (*b*) clarify the self-reported motivational structure behind educational-vocational decisions; and (*c*) examine the generality and stability of these factors.

Method

Instrument

To investigate the decision to attend college, an inventory, *Reasons for Going to College*, was developed through a series of pilot studies. Respondents were asked to show the degree of importance

[1] A version of this paper was presented to the Midwestern Psychological Association in Chicago, 1965. The research reported herein was supported by a grant from Cooperative Research Branch, United States Office of Education, Department of Health, Education and Welfare (Project 2109). It is part of a follow-up study on the determinants of educational choices. Joseph P. O'Reilly served as project assistant. The analysis was performed with the use of the facilities of the Statistical and Computing Center, University of Hawaii.

each of 78 reasons had in "influencing your decision to attend college." Each item was rated on a four-step scale ranging from "very important" to "of no importance." The opinionnaire items were assigned logically to one of the four groupings: Reasons, Values, Interests, and Influences. Most of the Reasons items were adapted from Iffert (1957); otherwise the items had been developed as part of a cross-sectional study of educational and vocational choices (Dole, 1961). Their concurrent validity and their reliabilities have been discussed elsewhere (Dole, 1961, 1965a).

Sample

The *Ss* included 214 male and 306 female students from the state university, who completed the inventory as freshmen and again as seniors. Since sex and college year were two variables of interest, the overall sample was broken down into four subsamples: freshman males, freshman females, senior males, and senior females. These students represented over 80% of all those who had survived 3 years at the University of Hawaii without delay.

Despite exotic connotations of the word Hawaii, Dole (1965a) has summarized evidence that University of Hawaii undergraduates are much like state-university populations across the country on standardized measures of personality, emotional stability, needs, interests, and abilities. Hawaii students are also comparable to students elsewhere in occupational goals and in father's occupational classification.

To contrast with the subsamples who had remained in college, a fifth group was also selected for study. This group included 350 male freshmen who subsequently had been dismissed because of low grades.

Administration of the Inventory

The inventory was distributed with freshman registration materials during the fall semester registration, 1960. Each return was audited individually. This procedure was repeated for the senior class in the fall of 1963. On the second administration the items were phrased retrospectively. That is, they were introduced with the additional phrase, "So far as you can remember now," and were expressed in the past tense.

Variables

Sixty-eight of the items from *Reasons for Going to College* were selected for analysis. (Ten items were discarded because more than 90% of the *Ss* in each of the four constant groups considered them of no importance to the college decision.) Total raw score on the Ohio State Psychological Examination (Form 25) and grade-point ratio [2] at the end of the first semester were also introduced as variables. It was assumed that there might be an academic-excellence factor among the inventory items and that these indexes of academic ability would be related to such a factor.

Factor Analyses

Product-moment correlations were completed among the 70 variables for the four subgroups, yielding four 70-order matrices. Examination of the latent roots of the full matrices (i.e., with unities in the main diagonal) revealed in each case 19 roots in excess of 1.00. Accordingly, it was concluded that 19 common factors were involved for all groups. Using squared multiple correlations as communality estimates, 19 principal axes factors were extracted and rotated by Saunders Equamax procedure.

Definition of Major Factors

For purposes of this study a major factor was defined as one which could be identified in all analyses when the factors resulting from the four analyses were matched. To match factors obtained in the four different analyses the following procedure was observed: (*a*) a factor was selected from the male freshman group, and items with loadings of .30 and above were noted; (*b*) a factor in the female freshman group was then identified which had a majority of items with loadings over .30 corresponding to the items in the freshman male group; (*c*) this procedure was continued for factors in the senior male and senior female groups; (*d*) finally, the factors in each group were reviewed for items with loadings from .15 to .29.

Whenever possible each of the 70 variables was assigned to but one major factor on the bases, first, that the variable loaded higher on that factor than on any other, and, second, that it had a loading on that factor of .30 or above in one freshman group or a loading of .15 or above in three out of four groups.

After one of the writers (Dole) had assigned all variables to major factors, the major factors were reviewed independently by the other (Digman) and discussed. Names were then assigned to the major factors on the basis of salient items, the results of other studies, and suggestions furnished by four other psychologists. Each group also exhibited a few factors specific to the group. These have been called "minor factors" and will not be discussed here.

Finally, as a check on the validity of the procedures and on the applicability of the major factors to other college freshman groups, the responses of the dismissed males were factor analyzed. By inspection, the 19 factors which were extracted were matched against the independently derived major factors.

[2] In computing grade-point ratio, A was assigned a value of 4, B of 3, C of 2, D of 1, and F of 0. Total honor points were divided by total credit hours for each student.

RESULTS

Factor Analyses

It was possible to distribute the 68 inventory variables among 13 major determinant factors.[3] In Table 1 the variables with their loadings have been grouped into these

[3] The results of these five factor analyses are presented in detail in Table G which has been

TABLE 1

FACTORS FROM THE FRESHMAN AND SENIOR SURVEYS

Factor	Loading				
	Dismissed	Freshman survey		Senior survey	
		Males	Females	Males	Females
Factor I—Social Reason					
Make new friends	49	66	67	66	51
Learn to get along with others	45	67	62	69	58
Develop socially	53	68	70	71	60
Contact advantageous	37	44	36	60	
Be more influential	25	41	46	38	17
Meet person to marry	36	36	35	45	28
Leadership civic affairs	19	46	38	34	
Factor II—Conformity					
Always expected	38	57	59		67
Person respected most	45	33	24	18	21
Friends going	39	55	41	47	28
Increase reputation	39	59	36	27	25
Family always gone	30	48		31	45
Decision made for me	16	64	65	52	69
Parents insisted		79	71	67	69
Girl like me expected to go	21		65		70
Fellow like me expected to go	21	68		75	
Value of prestige	24	43	23	19	
Influence of parents		79	71	67	69
Influence of friends	23		31	23	25
Influence of relatives	19	33	34	25	28
Factor III—Curiosity					
Serious curiosities	48	63[a]	42	28	37
Explore several lines			61	29	58
Find out about fields	42	47[a]	51	58	63
Factor IV—Vocational Reason					
Degree necessary for work	62	63	42[b]	64	30
Factor V—Academic Value					
Enjoyed studying	24	48	50	45	40
Degree for graduate school	18	53	27	52	16
Value of aptitude	47	26	22	49	
Value of specialization		35	35	33	25
Value of satisfaction	62	55	42	67	
Influence of high school courses	20	40	50	20	
Factor VI—Material Value					
Prepare self to be successful	16	28	44	44	18
Live an easier life	47	49	39	47	45
Degree meant great deal		35	44	40	
Raise my station in life	20	54	53	57	40
Value of practicality	42		25	42	44
Value of security	54	71	62	68	67
Value of potential income	67	82	72	80	75
Value of independence	46	52	18	45	44
Value of advancement	69	66	59	80	64
Value of easiest thing		17		30	33

TABLE 1—*Continued*

Factor	Loading				
	Dismissed	Freshman survey		Senior survey	
		Males	Females	Males	Females
Factor VII—Altruistic Value					
Value of serving others	43	37	60	32	55
Value of self-improvement	59	47	67	30	62
Value of parenthood	36	60	67	47	66
Value of religion	39	52	51	33	39
Interest in children and youth	18	28	28	27	31
Factor VIII—School Influence					
Community leaders encouraged		57	40	67	26
Teachers thought college material		52	60	53	65
Influence of high school teacher	52	22	57	31	57
Influence of school counselor	60		38	40	42
Influence of tests	30	16	46		56
Factor IX—Experience					
Influence of work experience		44	49		
Influence of movies and TV	46	40		52	40
Influence of people in field	40	68	51	31	47
Interest in work with adults	20	33		20	16
Factor X—Avocational Influence					
Influence of hobby	41	41[c]	57	48	54
Influence of free time	56	24[c]	33	60	54
Factor XI—Science Interest					
Influence of career day		32		29	
Interest in machines	53	73	57	69	31
Interest in numbers	66	67	51	71	61
Interest in science	22	27	55	50	72
Interest in plants and animals			51	17	64
Factor XII—Humanities Interest					
Interest in music	54	63	62	69	17
Interest in art	35	43	60	27	21
Interest in recreation	52	38	26	29	65
Interest in travel	38	30	32	27	44
Factor XIII—Verbal Interest					
Influence of reading	22	52	49	68	36
Interest in words	58	61	60	70	62
Interest in ideas	53	42	54	66	50
Factor XIV—Ability					
Ohio State Psychological	25	61	64	68	63
Grade-Point Average	54	63	62	54	64

Note.—Loadings below 15 not reported in this table. Decimal points omitted.
[a] Combines with Factor V in freshman constant males.
[b] Combines with Factor V in freshman constant females.
[c] Combines with Factor XI in freshman constant males.

major factors. The ability factor (XIV) included the Ohio State Psychological Examination and grade-point-ratio variables but none of the inventory items. Twelve minor factors are presented elsewhere (Dole, 1965a, Appendix II). In spite of its very high frequency of response it was decided to include the item, "I feel that a college degree is necessary for the kind of work I want to do." It is

deposited with the American Documentation Institute. Order Document No. 8521 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $2.50 for microfilm or $6.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

recognized that its emergence as a factor (IV) may have been an artifact.

For the analyses of the dismissed male group, item loadings are shown in the left-hand column of Table 1 wherever a fit with a major factor was identified. In the dismissed group 57 out of 68 inventory variables attained substantial loadings on 1 of the 13 major determinant factors and the ability factor again emerged.

DISCUSSION

Major Factors

Thirteen major factors have been identified in a questionnaire whose items ask students to rate the importance of various reasons for attending college. These factors were confirmed when measured among seniors in retrospect and among a group dropped for academic failure.

Reasons for College

These analyses also permit clarification of the motivations which students reported for their educational decision. On an a priori basis items had been classified as Reasons, Values, Interests, and Influences. Factor analysis justified a reclassification into 13 major factors.

A comparison may be fruitful between the factors identified in this research and those proposed by other behavioral scientists who have been concerned with the structure of educational and vocational motivation. Tyler[4] has called attention to the resemblance between the results reported here and her results (Tyler & Sundberg, 1964) with quite different methods and with different samples. Tyler and Sundberg were concerned with vocational rather than educational decisions. Tyler would interpret School Influence, Experience, and Avocational Influence (VIII, IX, and X) as choice strategies, and Social Reason, Conformity, and Vocational Reason (I, II, III) as values.

Super (1962) and O'Connor and Kinnane (1961) have discussed work values that appear similar to Social Reason, Academic Value, Material Value, and Altruistic Value. The authors' classification of Material and Altruistic Values is consistent with the research on college populations of Goldsen et al. (1960). Parallels are also apparent with Rosenberg's (1958) three value complexes—intrinsic reward (self-expressive), extrinsic reward (money, security), and people—or with Herzberg, Mausner, and Synderman's (1959) extrinsic and intrinsic motivations among executives. Bidwell et al. (1963) have proposed that Spindler's conception of a traditional-versus-emergent value conflict in the American culture be extended to the reported determinants of undergraduate careers. The Vocational Reason and Material Value factors seem similar to the traditional value category, and the Social Reason, Conformity, and Altruistic Value factors resemble their emergent value category.

It seems, however, that Tyler's stress on individuality and Super's work values are more congruent with the results of the present research than the simpler classifications proposed by Goldsen, Herzberg, Spindler, or Rosenberg. Two or three "value" orientations seem quite insufficient as explanatory concepts for complex human decisions.

Individual interviews with selected undergraduate students have reinforced the conception of educational decision as a result of three kinds of determinants—major factors, minor factors, and individual considerations. The major factors have been discussed above. Of the 12 minor factors which were identified in this study 3 were clearly common across three or more groups, 4 were sex related, and 5 were found in only one of the five samples. The present research has emphasized the major factors, principally because of their generality. Of course, additional, sometimes quite specific, factors will underlie the decision to attend college. The value of the "major factors" classification which has been presented here is that it furnishes a conceptual scheme within which, for example, a counselor with his individual case might operate, modifying and filling out where this seems necessary.

Generality of the Factors

Additional evidence that at least seven of the major factors in educational choice can be generalized to intermediate and high

[4] L. E. Tyler, personal communication, 1965.

school samples has been obtained in three additional analyses (Dole, 1965a): one analysis of 300 ninth-grade boys and girls from a Honolulu Intermediate school in a disadvantaged area; another of 300 college-bound pre-science ninth-grade males of Japanese ancestry; and a third consisting of 1,173 senior boys and girls from three high schools in Oakland, California. The samples were selected to emphasize diversity. The instrument, which included 39 items in checklist form adapted from the college inventory, appraised reasons for the choice of high school study program (college preparatory, technical, business, or terminal). Conformity, Academic Value, Material Value, Altruistic Value, School Influence, Experience, and Science Interest emerged rather clearly as major factors in these public school samples.

It was possible to match 32 items fairly closely with the college factors; 16 of them loaded significantly on similar factors in all eight factor analyses. Except for sex affiliation, the underlying factor structures were independent of various personal and social characteristics selected for analysis; they appeared in heterogeneous samples, including high school seniors from California; the items comprising the public school factors were scored by a different method (checklist) from the college factors (Likert) and presented in a different format and order; and the interrelationships were measured by phi rather than by product-moment correlation. It is felt then that at least seven major factors have considerable generality.

Stability

In addition to their generality with respect to other groups, there was interest in the stability of the factors in college attendance. When the college seniors retrospected—"So far as you can remember now"—about their motivations 3 years before, most of the same factors emerged. Of course, this tells little about the accuracy of individuals in recollecting, but it does strengthen confidence in the substantiality of the major factors within groups.

Order

In Table 1 the major factors were presented according to an a priori classification scheme. If the major factors are regrouped by inspection according to size of loadings for the variables, Material Value and Conformity generally have the highest loadings. Experience and Humanities Interest tend to have smaller loadings than the other factors.

CONCLUSIONS

Perhaps these factor analyses are sufficient to justify some speculations on the common motivational structures reported as a basis for educational decision. For many, college is primarily the path to a job and concomitant economic reward, while others are simply conforming to family or peer pressures. In contrast, there are young persons who find gratification in the academic enterprise for its own sake, enjoying intellectual inquiry or the pursuit of interests in art, letters, and science. Some are attracted by social opportunities; others heed the authority of the schools: "With your grades or test scores, of course you're going to college." Quite evidently the possible combinations of these factors are many and varied.

A logical next step would be to relate the 13 major factors to other variables. As reported elsewhere (Dole, 1965a), the factors extracted from *Reasons for Going to College* have been used as scales. These scales were treated as dependent variables in analyzing differences between males and females, between undergraduate college affiliations, and between those who persisted in college and those who later failed, transferred, or withdrew voluntarily. In general, the outcomes of these later studies have been informative and meaningful.

REFERENCES

BERDIE, R. F. *After high school what?* Minneapolis: University of Minnesota Press, 1954.

BIDWELL, C. E., KING, S. H., FINNIE, B., & SCARR, H. A. Undergraduate careers: Alternatives and determinants. *School Review*, 1963, 71, 299–316.

DOLE, A. A. *A study of values as determinants of educational-vocational choices in Hawaii.* Cooperative Research Project No. 757. Honolulu: Hawaii State Department of Education, 1961.

DOLE, A. A. Reported determinants of educational choice. *Personnel and Guidance Journal*, 1964, **42**, 564–571. (a)

DOLE, A. A. Sex as a factor in the determination of educational choice. *Journal of General Psychology*, 1964, **71**, 267–278. (b)

DOLE, A. A. *Follow-up studies of the determinants of educational-vocational choices*. Cooperative Research Project No. 2109. Washington: United States Department of Health, Education and Welfare, 1965. (a)

DOLE, A. A. Values as determinants of college attendance. Paper read at American Personnel and Guidance Association, Minneapolis, April 1965. (b)

FLANAGAN, J. C., DAVIS, F. B., DAILEY, J. T., SHAYCOFT, M. F., ORR, D. B., GOLDBERG, I., & NEYMAN, C. A. *The American high school student*. Cooperative Research Project No. 635. Pittsburgh: University of Pittsburgh, 1964.

GETZELS, J. W. Changing values challenge the schools. *School Review*, 1957, **45**, 92–102.

GOLDSEN, R. K., ROSENBERG, M., WILLIAMS, R. M., JR., & SUCHMAN, E. A. *What college students think*. New York: Van Nostrand, 1960.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. *The motivation to work*. New York: Wiley, 1959.

IFFERT, R. E. *Retention and withdrawal of college students*. Washington: United States Government Printing Office, 1957.

MCCLELLAND, D. C., BALDWIN, A. L., BRONFENBRENNER, U., & STRODTBECK, F. L. *Talent and society*. Princeton: Van Nostrand, 1958.

O'CONNOR, T. P., & KINNANE, T. F. A factor analysis of work values. *Journal of Counseling Psychology*, 1961, **8**, 263–267.

ROSENBERG, M. *Occupations and values*. Glencoe, Ill.: Free Press, 1958.

SANFORD, N. (Ed.) *The American college*. New York: Wiley, 1962.

SPINDLER, G. D. Education in a transforming American culture. *Harvard Educational Review*, 1955, **25**, 145–153.

SUPER, D. E. The structure of work values in relation to status, achievement, interests, and adjustment. *Journal of Applied Psychology*, 1962, **46**, 231–239.

TYLER, L. E., & SUNDBERG, N. D. *Factors affecting career choices of adolescents*. Cooperative Research Project, No. 2455. Eugene, Oregon: University of Oregon, 1964.

(Received June 17, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 254–265

DETERMINANTS OF SATISFACTION IN MIDDLE-MANAGEMENT PERSONNEL [1]

DILIP K. LAHIRI AND SURESH SRIVASTVA [2]

Indian Institute of Management, Calcutta, India

Responses to a questionnaire were obtained from 93 middle managers where Ss indicated the extent to which 13 job-content factors and 13 job-context factors contribute to the feeling of satisfaction and dissatisfaction in the present and the imaginary job situations. Results indicate that, contrary to the conventional assumption of bipolarity, satisfaction and dissatisfaction represent 2 distinct and separate feelings, and both groups of factors may act as satisfiers and as dissatisfiers. The present results were also compared with (a) the results obtained in a study of skilled workers, and (b) the American studies. The findings cast serious doubts on the generality of the Herzberg-Mausner-Snyderman motivator-hygiene theory.

This study was concerned with determining the underlying sources of satisfaction and dissatisfaction among employees occupying middle-management positions. Recent views on job satisfaction relate to two distinct and separate groups of elements that act differentially as sources of satisfaction and dissatisfaction: extrinsic elements in job environment which include physical and monetary rewards, and intrinsic elements which facilitate the employee self-actualization process.

Although the intrinsic and the extrinsic elements have not been empirically delineated as independent structures, Herzberg, Mausner, and Snyderman (1959) have used this dichotomy for relating job elements with satisfaction in their "motivator-hygiene" theory. Herzberg (1965b) in a recent article states:

> that job satisfaction and job dissatisfaction represent two separate and distinct experiences, and not just the opposites of the same feeling. What determines job dissatisfaction are those aspects of work which essentially describe the environment or surroundings within which one performs his work tasks. . . . Conversely, the elements of work which contribute to job satisfaction are those which essentially describe the relationship of the worker to what he does, his task, or job content as opposed to context [p. 369].

In essence, the motivator-hygiene theory suggests that the job-content factors or the "motivators," such as achievement, recognition, work itself, responsibility, advancement, and other matters associated with the self-actualization of the individual on the job, will produce satisfaction; however, lack of these factors does not produce dissatisfaction. On the other hand, the job-context factors or the "hygienes," such as salary, social and human relations aspects of the job, supervision, company policies and administrative practices, job security, working conditions, and several other factors somewhat peripheral to the task, tend to produce job dissatisfaction but their presence does not produce satisfaction. In effect, this theory suggests a nonlinear relationship between the job-content and the job-context factors, and satisfaction and dissatisfaction are not considered to be on a bipolar continuum. According to this theory, therefore, one would predict that managers will choose the job-content factors more often than the job-context factors while describing satisfying situations, and conversely, the job-context factors would be mentioned more often than the content-related factors as sources of dissatisfaction.

[1] The paper was read at the Section of Psychology and Educational Sciences, Indian Science Congress Session, Chandigarh, January 1966.

[2] Presently at the Alfred P. Sloan School of Management, Massachusetts Institute of Technology.

The authors are indebted to S. Sinha, Department of Psychology, Calcutta University; John Thomas, Indian Institute of Management, Calcutta; and S. P. Mukherjee, Department of Statistics, Calcutta University, for their valuable comments and criticisms of this study. Grateful acknowledgment is also made to D. K. Sanyal, Director, Indian Institute of Social Welfare and Business Management, Calcutta, for his permission in obtaining the sample of respondents from the Institute.

Herzberg et al. studies were followed by other studies which have attempted to test the generality of the above contentions (Centers & Bugental, 1966; Ewen, 1964; Friedlander, 1964; Halpern, 1966; Malinovsky & Barry, 1965; Myers, 1964; Saleh, 1964; Schwartz, Jenusaitis, & Stark, 1963; Wernimont, 1966). These results, in general, do not support the motivator-hygiene theory unequivocally.

For instance, Ewen (1964) found that for life insurance agents some of the satisfiers were related to satisfaction while others caused dissatisfaction and some of the dissatisfiers actually acted as satisfiers. Saleh (1964) reports that employees in the preretirement period preferred the job-context factors as the sources of satisfaction. Wernimont (1966) concludes from his study on engineers and accountants that both intrinsic and extrinsic factors can be sources of both satisfaction and dissatisfaction, but intrinsic factors are stronger.

Friedlander (1964) found very similar results. He found that intrinsic elements in the job situation were important for both job satisfaction and job dissatisfaction, while extrinsic aspects were relatively unimportant as satisfiers or as dissatisfiers. Friedlander (1963) and Malinovsky and Barry (1965), using the factor-analytic technique, have found that both motivator and hygiene variables were associated with job satisfaction. Factors that emerged in these studies correspond, in part, with Herzberg's concept of motivator and hygiene variables, while other factors were found to include both motivator- and hygiene-type variables. Choudhuri and Lahiri (1966) found that for skilled workers content-related factors and context-related factors may be both sources of satisfaction and dissatisfaction, but context-related factors are more important as satisfiers than the content-related factors. Centers and Bugental (1966) found that intrinsic and extrinsic job components were valued differentially at different occupational levels. For instance, the white-collar population (professional and managerial, clerical, and sales personnel) placed a greater value on intrinsic sources of job satisfaction. The blue-collar workers (skilled, semiskilled, and unskilled), on the other hand, placed a greater value on extrinsic sources of job satisfaction.

A number of writers have criticized the methods used by Herzberg et al. (Brayfield, 1960; Ewen, 1964; Kahn, 1961; Vroom, 1964; Vroom & Maier, 1961). These authors have suggested that Herzberg et al. results, apart from the weakness of content analysis of critical incidents reported during interviewing situations, may be interpreted in an entirely different way. It is possible that stated sources of satisfaction may result from the general tendency to attribute the causes of satisfaction to one's own success and achievements in the job. While in some cases sources of dissatisfaction may stem not from personal inabilities, as one may think, but from external factors in the work situation, that is, mangagement policies or supervisor, people generally tend to attribute the causes of all failures to external objects and not to personal incapacities. Thus, sources of satisfaction and dissatisfaction may be an expression of defensive forces within an individual.

In view of the lack of conclusive evidence projected in the above-mentioned studies, it may be said that the role of the job-content factors and the job-context factors as sources of satisfastion and dissatisfaction is not clearly established. The purpose of this study is to replicate and test the motivator-hygiene theory in a different population and in a different culture. Specifically, this study is concerned with the analysis of job elements which contribute to the feeling of satisfaction and dissatisfaction in middle-management personnel.

Four hypotheses have been tested in this investigation. The first hypothesis is related to the bipolarity of the satisfaction–dissatisfaction dimension. The question is: do job satisfaction and job dissatisfaction represent two separate and distinct feelings and not a bipolar dimension on a single continuum? The other hypotheses are related to the question: do the job-content factors and the job-context factors act differentially in producing satisfaction and dissatisfaction? Do the job-content factors act as satisfiers and the job-content factors act as dissatisfiers? The following hypotheses were tested:

Hypothesis 1

If a job factor or the positive aspect of it in a job situation contributes to the feeling of employee satisfaction, the lack of or the negative aspect of that factor will not necessarily contribute to the feeling of employee dissatisfaction in the same degree. *Conversely,* if the lack of a job factor or the negative aspect of it in a job situation contributes to the feeling of employee dissatisfaction, the presence of or the positive aspect of that factor will not necessarily contribute to the feeling of employee satisfaction in the same degree.

Hypothesis 2

The positive aspect of the job-content factors (or motivators) in a job situation will tend to contribute to the feeling of employee satisfaction *more* than the positive aspect of the job-context factors (or hygienes) in that job situation. *Conversely,* the positive aspect of the job-context factors (or hygienes) in a job situation will tend to contribute to the feeling of employee satisfaction *less* than the positive aspect of the job-content factors (or motivators) in that job situation.

Hypothesis 3

The lack of the job-context factors (or hygienes) or the negative aspect of them will tend to contribute to the feeling of employee dissatisfaction *more* than the lack of the job-content factors (or motivators) or the negative aspect of them in that job situation. *Conversely,* the lack of the job-content factors (or motivators) or the negative aspect of them in a job situation will tend to contribute to the feeling of employee dissatisfaction *less* than the lack of the job-context factors (or hygienes) or the negative aspect of them in that job situation.

Hypothesis 4

The positive aspect of the job-content factors (or motivators) in a job situation will tend to contribute to the feeling of employee satisfaction, but the lack or the negative aspect of those factors in that job situation will not tend to contribute to the feeling of employee dissatisfaction; and even if it does, it would be to a lesser degree. *Conversely,* the lack or the negative aspect of the job-context factors (or hygienes) in a job situation will tend to contribute to the feeling of employee dissatisfaction, but the positive aspect of those factors in that job situation will not tend to contribute to the feeling of employee satisfaction; and even if it does, it would be to a lesser degree.

In addition, two other questions, peripheral to the principal hypotheses, were examined: (*a*) Are the satisfiers and the dissatisfiers in different job levels identical? Are the satisfiers and the dissatisfiers in lower job levels (skilled workers) and high job levels (middle management) similar? (*b*) Are the two groups of factors found cross-culturally?

METHOD

Sample

Respondents were 93 executives holding middle-management positions from a variety of industries. The subjects (*Ss*) were attending executive development programs at the Indian Institute of Management, Calcutta, and evening courses in business management at the Indian Institute of Social Welfare and Business Management, Calcutta. The *Ss* ranged in age from 20 to 50 yr. and had a median age of 33 yr. The range of total job experience was from 5 to 25 yr. ($Mdn = 7.8$ yr.) and total work experience on the same position was from 2 to 15 yr. ($Mdn = 4.6$ yr.). In terms of education, 67% were graduates and 33% had masters degrees. All were males and most were married (85%), and belonged to middle-class educated families (88%).

Questionnaire and Procedure

The data for this study were obtained by means of a questionnaire. This questionnaire contained four parts. The first part consisted of items concerning the demographic characteristics of the respondents. The other parts of the questionnaire were related to the measures of satisfaction and dissatisfaction with different job factors. The job factors chosen for this study included 13 satisfiers and 13 dissatisfiers. These 26 factors were selected from different literature (Friedlander, 1964; Herzberg et al., 1959; Sinha, 1958) and from the respondents in a short pilot study. Four scales were used for measuring satisfaction and dissatisfaction. Two scales were used to measure satisfaction and dissatisfaction with the present job, while two other scales were developed to measure satisfaction and dissatisfaction with an imaginary job.

Directions for the first of these measures were as follows:

> Think of a time when you felt exceptionally good about your present job either recently or any other time. The following are some of the factors which may have contributed to your feeling of satisfaction at that time. Indicate by checking (√) the extent to which each factor contributed to your feeling of satisfaction on the scales given below.

Directions for the second measure were the same except the word "bad" and "dissatisfaction" were substituted for "good" and "satisfaction." In addition to these two scales asking for ratings of *S*'s present job, the other two scales were related to imaginary job situations. The first asked *Ss* to indicate the extent to which the lack of or the negative aspect of each factor might contribute to the feeling of dissatisfaction in a "worst" job he could imagine himself doing, and, the second, the extent to which the same factor might contribute to the feeling of satisfaction in a "best" job he could imagine himself doing.

TABLE 1

INTRACLASS CORRELATION COEFFICIENTS FOR MOTIVATORS, HYGIENES, AND TOTAL JOB VARIABLES

Measure	Present job			Imaginary job		
	Motivators	Hygiene	Total	Motivators	Hygiene	Total
Satisfaction	.80*	.79*	.75*	.51*	.64*	.76*
Dissatisfaction	.45*	.74*	.53*	.55*	.79*	.78*

* $p < .01$.

The satisfaction and the dissatisfaction measures were identical. In both the satisfaction and the dissatisfaction measures there were included a series of horizontal 10-point graphic scales with the end points labeled "no satisfaction or no dissatisfaction" on one side and "maximum satisfaction or maximum dissatisfaction" on the other side. The *Ss* were asked to mark the point along the line which showed the contribution ascribed by the respondents to each of the 26 factors as sources of satisfaction and the lack of or the negative aspect of the same 26 variables as sources of dissatisfaction. Thus, for each job factor four measures were obtained, two for satisfaction and two for dissatisfaction; two for the present job and two for the imaginary job. To eliminate the possible "response set," job factors were randomly presented in each of the four scales. This dyadic (the present job and the imaginary job) scaling technique has at least one advantage over the other methods (Friedlander, 1964; Herzberg et al., 1959) in that it takes into account the frame of reference of the individual instead of asking his entire past vocational experience. An a priori assumption has been made here that satisfiers and dissatisfiers in different job levels are different. A job factor which acts as a satisfier or a dissatisfier in one job may lose entirely its significance in its role as a satisfier or as a dissatisfier in another job. But the imaginary job might provide *S* with an opportunity to respond to job factors without being bound by the salient job factors of either the present or the other jobs he might have held previously. In some sense the satisfying or dissatisfying job factors will become more significant than the particular experience at a job.

RESULTS

Reliability

The reliabilities of the satisfaction and the dissatisfaction measures for motivators, hygienes, and total job variables were computed separately by means of the Ebel's intraclass correlation formula (Guilford, 1954, p. 395) for mean ratings from K raters. The coefficients obtained from the data are presented in Table 1 and were tested for significance ($p < .01$). It can be seen that these reliability estimates were sufficiently high, indicating a degree of consistency among the respondents with respect to the motivators and the hygienes as sources of satisfaction and dissatisfaction.

Testing of Hypothesis

Table 2 gives several statistics for scores in satisfying and dissatisfying situations under the present and the imaginary job situations for each of the job factors belonging to the motivators and the hygienes. Means and standard deviations of satisfaction and dissatisfaction scores, respectively, in the present job appear in Columns 1–4 and corresponding figures for the imaginary job in Columns 5–8. Mean difference (satisfaction–dissatisfaction) scores for the present and the imaginary jobs are given in the last two columns. As indicated by the asterisks, for 20 of the 26 factors (mostly motivators) the mean satisfaction measure was significantly higher than the mean dissatisfaction measure. This implies that the presence of any of these factors gives more satisfaction than the amount of dissatisfaction caused by its absence. In other words, dissatisfaction is not the reverse of satisfaction for these factors and the same continuum cannot be advocated for measures representing these two feelings. However, mean amounts of satisfaction and of dissatisfaction for the remaining 6 job factors, namely, supervisor's help, salary, fairness of authority, freedom of expression, recognition, and growth, are not significantly different among themselves.

The overall results are displayed by the two-way analysis of variance presented in Table 3. As expected, mean satisfaction and dissatisfaction measures for all the factors

TABLE 2

MEANS AND STANDARD DEVIATIONS BETWEEN SATISFACTION AND DISSATISFACTION MEASURES FOR THE PRESENT JOB AND FOR THE IMAGINARY JOB SITUATIONS

Factor	Present job (P)				Imaginary job (I)				Difference	
	Satisfaction		Dissatisfaction		Satisfaction		Dissatisfaction			
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	P	I
Hygienes										
Relations with co-workers	6.82	2.28	3.82	2.90	8.09	1.98	5.29	3.05	3.00*	2.80*
Superior's help	6.12	2.82	5.59	3.25	7.41	2.62	6.38	2.97	0.53	1.03
Friendliness of superior	7.15	2.86	5.26	3.25	7.12	3.07	6.09	3.26	1.89*	1.03
Technical competence of superior	5.44	2.81	5.12	3.47	7.47	2.65	5.94	2.80	.32	1.53*
Salary	4.94	2.79	6.06	3.21	7.38	2.80	7.23	2.82	−1.12	.15
Security	7.29	2.95	4.53	3.27	8.29	2.41	5.50	3.41	2.76*	2.79*
Working conditions	6.53	3.06	3.24	2.45	7.62	2.48	5.38	3.01	3.29*	2.24*
Benefits	6.47	2.91	4.15	2.97	7.53	2.66	5.82	3.06	2.32*	1.71*
Fairness of authority	6.12	3.19	4.70	3.21	7.91	2.57	7.26	2.96	1.42	.65
Freedom of expression	5.79	3.03	4.82	3.34	7.76	2.64	7.03	2.69	.97	.74
Work group	5.12	2.54	4.38	2.61	7.03	2.47	5.53	2.39	.74	1.50*
Managerial policies	5.12	2.67	6.73	2.88	6.59	2.96	6.82	3.16	−1.61*	−.23
Home life	8.00	2.53	3.65	3.02	8.82	2.41	3.82	3.46	4.35*	5.00*
Motivators										
Recognition	6.32	2.89	5.00	3.15	7.97	2.50	6.91	3.01	1.32	1.06
Challenging assignments	6.59	3.14	4.47	3.30	7.73	2.49	5.56	3.45	2.12*	2.17*
Growth	6.09	3.00	5.26	3.25	7.88	2.31	7.06	2.91	.83	.82
Achievement	7.03	2.75	4.82	3.12	8.68	2.07	6.06	3.14	2.21*	2.62*
Liking for the work	7.00	2.41	4.68	3.13	8.76	1.87	6.12	3.37	2.32*	2.64*
Accomplishment	7.65	1.99	4.44	2.96	8.21	2.10	6.38	2.89	3.21*	1.83*
Use of best abilities	7.18	2.43	4.38	3.23	8.53	2.30	6.59	2.85	2.80*	1.94*
Responsibility	8.00	2.11	4.24	3.53	8.68	1.89	5.65	3.13	3.76*	3.03*
Autonomy	6.35	2.57	4.50	2.99	7.29	2.77	7.41	2.42	1.83*	−.12
Promotion	4.59	2.99	6.18	3.17	7.53	2.80	7.18	2.79	−1.59	.35
Prestige	6.91	2.85	5.12	3.53	8.18	2.24	7.35	2.74	1.79*	.83
Work itself	6.35	2.92	4.61	3.18	8.53	2.01	6.44	3.26	1.74*	2.09*
Status	6.73	2.67	5.18	3.11	8.26	2.44	6.65	2.93	1.55*	1.61*

* $p < .05$.

TABLE 3

ANALYSIS OF VARIANCE OF SATISFACTION-DISSATISFACTION MEASURES AND MOTIVATOR-HYGIENE VARIABLES FOR THE PRESENT JOB AND FOR THE IMAGINARY JOB SITUATIONS

Factor	*df*	Present job		Imaginary job	
		MS	*F*	*MS*	*F*
Satisfaction-dissatisfaction scale (S)	1	1195.77	71.47*	1200.71	97.70*
Motivators-hygiene variables (V)	1	29.40	1.76	30.18	2.46
S × V	1	16.73	1.77	12.29	1.24
Within cells	4832	9.43		9.93	

* $p < .01$.

taken together differ significantly, thus supporting the first hypothesis that job satisfaction and job dissatisfaction are two separate and distinct feelings and not the opposites of the same continuum. However, differences in the mean scores for the two groups (motivators and hygienes) do not come out to be significant and, naturally enough, differential behavior of satisfaction–dissatisfaction scales with groups turns out to be insignificant also. This insignificant difference may be explained in the following terms: the means of the motivators in describing the satisfying situation are greater than the means of the hygienes in that situation, and the means of the

TABLE 4

COMPARISON OF MEAN SCORES BETWEEN THE GROUPS' MOTIVATORS AND HYGIENES

Job situation	Scale	Average difference score ($\bar{d}$)	t	Comment ($\alpha = .01$)
Present job	Satisfaction	5.59	1.92	Difference between mean satisfaction scores for motivators and hygienes is insignificant.
Present job	Dissatisfaction	0.82	0.26	Difference between mean dissatisfaction scores for motivators and hygienes is insignificant.
Imaginary job	Satisfaction	6.56	3.75	Difference between mean satisfaction scores for motivators and hygienes is significant.
Imaginary job	Dissatisfaction	4.74	1.40	Difference between mean dissatisfaction scores for motivators and hygienes is insignificant.

motivators in describing the dissatisfying situation are less than the means of the hygienes in that situation, as significant differences in opposite directions on the satisfaction and the dissatisfaction scales separately may give rise to average differences apparently attributable to random variations above.

To look into this, paired t comparisons for each of the cases noted above were carried out. The results are presented in Tables 4 and 5. It is worth noticing in Table 4 that under the present job situation the mean satisfaction or dissatisfaction measures for "motivators" and "hygienes" are not significantly different. It was also observed that in all cases the mean satisfaction scores and mean dissatisfaction scores for motivators were higher than those for hygienes. As shown in Table 5, among the motivators as well as among the hygienes, the degree of satisfaction expressed exceeds significantly the degree of dissatisfaction. The results thus support Hypothesis 4 that the presence of motivators causes satisfaction to a greater extent than the dissatisfaction caused by their absence. However, the lack or the negative aspect of motivators also contributes to the feeling of employee dissatisfaction as shown in the result that the mean dissatisfaction score for motivators differs significantly from zero. In other words, motivators also act as dissatisfiers but to a lesser extent than as satisfiers.

With repect to hygienes, the hypothesis that they act more as dissatisfiers than as satisfiers does not stand test. The lack or the negative aspect of hygienes does not contribute to the feeling of employee dissatisfaction more than the amount of satisfaction given by its positive aspect. In fact, the result shows that except possibly for four factors (viz., superior's help, salary, fairness of authority, and freedom of expression) hygienes, just like motivators, act more as satisfiers than as dissatisfiers.

TABLE 5

COMPARISON OF MEAN SCORES BETWEEN THE SCALES' SATISFACTION AND DISSATISFACTION

Job situation	Scale	Average difference score ($\bar{d}$)	t	Comment ($\alpha = .01$)
Present job	Motivators	23.94	3.29	Mean satisfaction score is significantly greater than mean dissatisfaction score.
Present job	Hygienes	19.20	3.41	Mean satisfaction score is significantly greater than mean dissatisfaction score.
Imaginary job	Motivators	22.79	4.23	Mean satisfaction score is significantly greater than mean dissatisfaction score.
Imaginary job	Hygienes	21.03	3.66	Mean satisfaction score is significantly greater than mean dissatisfaction score.

TABLE 6

COMPARISON OF RANKINGS OF JOB FACTORS AS A SOURCE OF SATISFACTION AND AS A SOURCE OF DISSATISFACTION BY WORKERS AND MIDDLE-MANAGEMENT PERSONNEL

Factor	Satisfaction		Dissatisfaction	
	Workers	Middle management	Workers	Middle management
Benefits (welfares)	11.5 (13.5)	6 (15)	1 (1)	10 (24)
Salary	11.5 (13.5)	11 (26)	2 (2)	3 (3)
Promotion	10 (12)	12 (27)	3 (3)	2 (2)
Working conditions	7 (9)	15 (14)	4 (4)	12 (27)
Management policies	8 (10)	10 (24.5)	5 (5)	1 (1)
Recognition	9 (11)	9 (18)	6 (7)	5 (10)
Relations with supervisors	5 (5)	3 (6)	7 (9)	4 (5.5)
Autonomy	6 (6)	7.5 (16.5)	8 (10)	8 (18)
Work itself	4 (4)	7.5 (16.5)	9 (11)	6 (15)
Responsibility	2.5 (2.5)	1 (1.5)	10 (12)	9 (23)
Relations with co-workers	2.5 (2.5)	4 (10.5)	11 (13)	11 (25)
Security	1 (1)	2 (4)	12 (14)	7 (17)
rho	.77*		.26	

Note.—Low numbers represent high satisfaction or dissatisfaction and high numbers low satisfaction or dissatisfaction. The original rankings are in parentheses.
* $p < .01$.

Comparison with Workers' Data

In Table 6 is presented ranking obtained in a previous study of skilled workers for factors included in the present study. The number of overlapping factors between these two studies was 12. The ranks reported for these two studies are not consecutive since only 12 overlapping factors are presented in Table 6. For the purpose of comparison, the overlapping factors were reranked and rank-order correlation coefficients (*rho*) were computed and tested for significance. The values of *rho* appear at the bottom of Table 6. The comparison shows significant correlation ($p < .01$) between ranking of job factors as sources of satisfaction and insignificant correlation between ranking of job factors as sources of dissatisfaction. The results thus show that the sources of satisfaction for the two groups are common while job dissatisfaction stems from entirely different sources.

Comparison with American Data

In Table 7 and Table 8 are presented ranking of job factors as sources of satisfaction and as sources of dissatisfaction obtained in the American studies. The present data on each job item cannot be compared directly with other data since the methods were different in different studies. Rank-order techniques were used for comparison of other data. The number of overlapping factors in these studies varied from 10 to 15. The ranks reported in Friedlander's (1964) study are not consecutive, indicating that this study included one factor not contained in the present study. The overlapping factors in the study and each of the others were reranked in each case and rank-order correlation coefficients (*rho*) were computed and tested for significance. The values of *rho* appear at the bottom of Table 7 and Table 8.

The comparisons show low and insignificant correlations with the present study and the American studies in both satisfying and dissatisfying situations. For the American studies the correlations were higher and significant in satisfying situations, and higher and significant in 5 of the 10 cases of dissatisfying situations.

The low correlations between the present study and the American studies indicate that Indian workers and American workers placed different importance on individual job elements as sources of satisfaction and as sources of dissatisfaction. The higher correlations within the American studies in describing satisfying situations seem to result from nearly the same relative importance placed by American workers on individual job factors as satisfiers. The failure of the considerably higher correlation within the American studies in describing dissatisfying situations suggests that the converse is not necessarily true. This means that the American workers included in these studies did not place the same relative importance on individual job factors in describing dissatisfying situations.

DISCUSSION

The results of this study partly confirm and partly reject the studies mentioned previously. This study confirms Herzberg et al. (1959) and Friedlander's (1964) findings that satisfaction and dissatisfaction are not the opposite poles of the same feeling. The presence and the absence or the negative aspect of a job factor do not contribute to equal but opposite feelings. Employees who perceived

TABLE 7

COMPARISON OR RANKING OF JOB FACTORS AS SOURCES OF SATISFACTION BY INDIAN AND AMERICAN WORKERS

Job factor	Present study (1966)	Herzberg et al. (1959)	Schwartz et al. (1963)	Friedlander (1964)	Wernimont (1966)	
	Middle management ($N = 93$)	Engineers and Accountants ($N = 203$)	Supervisors ($N = 111$)	Miscellaneous ($N = 80$)	Engineers ($N = 82$)	Accountants ($N = 50$)
Hygienes						
Relation with co-workers	10.5	11	10	9	5	6
Superior's help	19.5	—	—	—	—	—
Friendliness of (Interpersonal relations) superior	6	8.5	5	6	—	—
Technical competence of superior	23	11	12	10.5	10	10
Salary	26	6	8	—	9	8
Security	4	14	14.5	10.5	—	—
Working conditions	14	14	14.5	12	8	7
Benefits	15	—	—	17	—	—
Fairness of authority	19.5	—	—	—	—	—
Freedom of expression	22	—	—	—	—	—
Work group	24.5	—	—	13	—	—
Management policies	24.5	11	9	15	7	9
Home life (Personal life)	1.5	14	12	18	—	—
Motivators						
Recognition	18	2	2	4	4	5
Challenging assignments	13	7	7	5	—	—
Growth	21	—	—	2.5	—	—
Achievement	7	1	1	2.5	1	1
Liking for the work	8	—	—	—	—	—
Accomplishment	3	—	—	—	—	—
Use of best abilities	5	—	—	7	—	—
Responsibility	1.5	4	3	8	3	2
Autonomy	16.5	—	—	—	—	—
Promotion (advancement)	27	5	4	14	6	4
Prestige	9	—	—	—	—	—
Work itself	16.5	3	6	1	2	3
Status	12	8.5	12	—	—	—
rho		−0.19	−.10	.14	.61	.56
			.91*	.80**	.83**	.82**
				.59**	.76**	.78**
					.82**	.67
						.92**

* $p < .05$.
** $p < .01$.

certain factors of work environment as sources of satisfaction may not perceive the absence or the negative aspect of the same factors as sources of dissatisfaction. For job factors like promotion, employee benefits, home life, management policies, and use of best abilities in Friedlander's (1964) study, and for job elements like superior's help, security, freedom of expression, recognition, and growth in the present study, the assumption of some bipolarity of the satisfaction-dissatisfaction continuum has been partially substantiated. Since the job factors in these two studies do not overlap, the questions of bipolarity with

TABLE 8

COMPARISON OF RANKING OF JOB FACTORS AS SOURCES OF DISSATISFACTION BY INDIAN AND AMERICAN WORKERS

Job factor	Present study (1966)	Herzberg et al. (1959)	Schwartz et al. (1963)	Friedlander (1964)	Wernimont (1966)	
	Middle management ($N = 93$)	Engineers and Accountants ($N = 203$)	Supervisors ($N = 111$)	Miscellaneous ($N = 80$)	Engineers ($N = 82$)	Accountants ($N = 50$)
Hygienes						
Relations with co-workers	25	9.5	13.5	17	9	9
Superior's help	4	—	—	—	—	—
Friendliness of superior	5.5	5	4	8	—	—
Technical competence of superior	8.5	2	5	12	6	6
Salary	3	4	10	—	7	8
Security	17	15	10	13	—	—
Working conditions	27	7.5	7	15.5	10	10
Employee benefits	24	—	—	18	—	—
Fairness of authority	13	—	—	—	—	—
Freedom of expression	11.5	—	—	—	—	—
Work group	21.5	—	—	11	—	—
Management policies (Company policies)	1	1	1	6	8	7
Home life	26	12.5	13.5	14	—	—
Motivators						
Recognition	10	3	2	4	1	2
Challenging assignments	19	—	—	3	—	—
Growth	5.5	9.5	15	7	—	—
Achievement	11.5	11	3	2	4	4
Liking for the work	14	—	—	—	—	—
Accomplishment	20	—	—	—	—	—
Use of best abilities	21.5	—	—	1	—	—
Responsibility	23	12.5	10	9	3	3
Autonomy	18	—	—	—	—	—
Promotion (Advancement)	2	7.5	8	10	2	1
Prestige	8.5	—	—	—	—	—
Work itself	15	6	6	5	5	5
Status	7	14	12	—	—	—
rho		.50	.30	.55	.29	.33
			.64*	−.33	.72*	.78*
				.62*	.55	.62
					.24	.27
						.98*

* $p < .01$.

respect to these job factors can neither be accepted nor denied.

The results of this study conflict with the Herzberg et al. construct of the basic two-dimensional nature of job attitude. In the Herzberg et al. conceptual framework one would expect that the intrinsic job factors will be endorsed as satisfiers while environmental factors are endorsed as dissatisfiers. The results of the present study show only partially that the Herzberg et al. (1959) framework is workable; both intrinsic and extrinsic job factors were found to be determinants of satisfied and dissatisfied feelings

toward the job. Intrinsic elements, however, act more as satisfiers and extrinsic elements act more as dissatisfiers. These findings are more consistent with results obtained by Ewen (1964), Friedlander (1964), and Wernimont (1966).

It may be seen from the results that responsibility, accomplishment, use of best abilities, achievement, liking for the job, security, friendliness of superior, relations with co-workers, and home life ranked very high on the list while describing a satisfying situation; in describing dissatisfying situations, promotion, prestige, growth, status, recognition, management policies, salary, fairness of authority, superior's help, and friendliness of superior ranked very high on the list. It is evident from these results that the respondents in this study endorsed job factors differently from what the motivator-hygiene theory would have predicted.

The factor of interpersonal relations with the superiors and the co-workers had been endorsed by the *Ss* of the present study as a contributor to job satisfaction more often than other extrinsic items when describing satisfying situations. Social and human relations aspects of the job as contributors to satisfaction have been supported by the results obtained by Schwartz, Jenusaitis, and Stark (1963), Friedlander (1963, 1964), and Wernimont (1966). In the results obtained by Russian social scientists as reported by Herzberg (1965a), the factor of relationship among fellow workers was found to be least discriminating between satisfied and dissatisfied workers. Herzberg (1965a) noted that this factor "was least discriminating because both satisfied and dissatisfied workers in Russian industry rated themselves as highly content [p. 250]" with this factor. In view of the conformity among these results with respect to this factor, Herzberg's (1965b) interpretation that "these inversions (crossing over of a hygiene factor to the motivator side), represent the pathology of job attitude feelings [p. 372]" may not be accepted as such.

When the results of the present study are compared with the results obtained from another study on skilled factory workers, the comparison shows that satisfaction may come from the intrinsic and extrinsic elements in both groups; however, dissatisfaction results from different sources. In both of the groups, extrinsic and intrinsic elements may act as satisfiers and dissatisfiers; for instance, for workers the extrinsic elements are more important, while for middle management the intrinsic elements play a major role.

The low and insignificant correlations between the present study and the American studies suggest that relative importance placed by the Indian and the American workers on individual job items are different. Such differences between the Indian and the American results are not altogether unexpected and reflect cultural and social differences between the two countries. However, in interpreting such data one should be cautious since the insignificant correlations might result from the different techniques employed by different studies.

The results obtained in this study may be explained in several possible ways. In contrast to the Herzberg et al. prediction, Malinovsky and Barry (1965) found a positive correlation (.55) between Factor II hygienes and overall job satisfaction. The positive correlations between motivators and overall job satisfaction, and hygienes and over-all job satisfaction, suggest that both motivators and hygienes were positively related to job satisfaction. The results of the present study thus are clearly in tune with other studies (Friedlander, 1963, 1964; Malinovsky & Barry, 1965; Wernimont, 1966). All these studies indicate that the motivator-hygiene theory implies a rather too rigid classification of job factors which contribute to satisfaction and dissatisfaction feelings of the employees.

When the present study was compared with the results obtained from a study on skilled workers, it was found that the two groups put different emphasis on different job factors as satisfiers and as dissatisfiers. In such a comparison, the occupational variable, it seems, comes into play. In general, blue-collar workers had low salaries, low education, and low social status. Probably, to these workers, the need underlying the hygienes becomes salient. Therefore, blue-collar workers are more concerned and preoccupied with fulfilling the prepotent needs (security, salary,

good working conditions, etc.) than workers in higher job levels. Once these needs are satisfied the higher-order needs related to self-actualizing aspects of the job would emerge, whereas for workers in higher job levels who had usually higher education and higher social status, these prepotent needs are basically fulfilled, and for them the high-order needs become more poignant and important. Therefore, the middle-management personnel found that the intrinsic factors of the job act more as satisfiers than do the extrinsic job factors. That occupational level is a meaningful variable in predicting need fulfillment has been demonstrated in Porter's study of job attitudes in management. Porter (1962, 1963) found that vertical level of position within management had a strong relation to the degree of perceived need satisfaction and perceived importance of needs. Porter and Lawler (1965) from their extensive review conclude that level shows a definite relationship to the employees' perceived job and need satisfaction. This explanation stems from Maslow's (1954) hierarchy of needs. From this point of view, blue-collar workers are not able to gratify the self-actualization need as represented by motivators since the needs that must be gratified first for this group are the more prepotent needs, that is, the safety need which is the basis of security, and the need for affection and love which underlie social and human relations aspects of the job. Hence, the extrinsic job factors act more as satisfiers and as dissatisfiers than intrinsic elements in this group. In higher job levels the prepotent needs are gratified (at least at a certain level) and, therefore, the motivators act more as satisfiers than the hygienes.

From Maslow's need hierarchy theory one would expect that middle-management personnel would find motivators as the only sources of satisfaction. The present results and the results of the other investigators have shown that for workers in higher job levels extrinsic elements also act as satisfiers. This may be explained in the following terms. If the employees, under the circumstances, find that they have little chance to gratify the needs underlying the motivators, they would choose more probably the hygiene factors thinking that these factors can provide at least some satisfaction. The assumption in this explanation is that the prepotent needs have been gratified in middle management, therefore, the employees in these jobs will try to gratify the higher need, that is, self-actualization. But if they perceive that the chances to gratify this need through the motivators are not immediately available, the middle-management personnel will seek other substitute goals despite the fact that they do not provide true satisfaction.

Another explanation could be found in the theory of cognitive dissonance. According to this theory (Festinger, 1957: Festinger & Aronson, 1960) one would experience a dissonance if one fails to reach a goal for which one is striving. And, dissonance, thus produced, "being psychologically uncomfortable, will motivate the person to try to reduce the dissonance and achieve consonance [Festinger, 1957, p. 3]." One way of reducing the dissonance is to shift the inner attitudes so that it stands more realistically to the external environment and to find some surrogates in the situation to which one can attach value. The cognition a middle manager has that the job situation is unpleasant since the job situation is not conducive to fulfill his high-order needs (motivators) is dissonant with his cognition that he must work in that work situation. He may reduce his dissonance, at least in one way, by convincing himself that the situation is not as unpleasant as it first appeared.

In such circumstances the middle managers would try to find many "good aspects" in the environmental factors of their job and derive satisfaction from them when they would find that the motivators become hard to attain despite any expanded effort. To reduce dissonance aroused from the lack of advancement or recognition, or other motivators, the middle managers choose the factors available—the hygienes, while describing a satisfying situation.

It seems that two types of interpretations may hold good, and satisfaction and dissatisfaction from job elements may stem from entirely different reasons than those which the motivator-hygiene theory would predict.

To sum up, the present study of middle-

management personnel shows that satisfaction and dissatisfaction are not on a bipolar continuum and both motivators and hygienes can be the sources of satisfaction and dissatisfaction. The motivator-hygiene theory, thus, is not fully supported in the present study.

REFERENCES

Brayfield, A. H. Review of Herzberg, F., Mausner, B., and Snyderman, B. The motivation to work. *Personnel Psychology*, 1960, **13**, 101–102.

Centers, R., & Bugental, D. E. Intrinsic and extrinsic job motivations among different segments of the working population. *Journal of Applied Psychology*, 1966, **50**, 193–197.

Choudhuri, P. K., & Lahiri, D. K. Perceived job characteristics as satisfiers and dissatisfiers by manual workers. Paper read at Indian Science Congress, Chandigarh, January 1966.

Ewen, R. B. Some determinants of job satisfaction: A study of the generality of Herzberg's theory. *Journal of Applied Psychology*, 1964, **48**, 161–163.

Festinger, L. *A theory of cognitive dissonance.* Evanston, Ill.: Row, Peterson, 1957.

Festinger, L., & Aronson, E. The arousal and reduction of dissonance in social contexts. In D. Cartwright & A. Zander (Eds.), *Group dynamics.* Evanston, Ill.: Row, Peterson, 1960.

Friedlander, F. Underlying sources of job satisfaction. *Journal of Applied Psychology*, 1963, **47**, 246–250.

Friedlander, F. Job characteristics as satisfiers and dissatisfiers. *Journal of Applied Psychology*, 1964, **48**, 388–392.

Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

Halpern, G. Relative contributions of motivators and hygiene factors to overall job satisfaction. *Journal of Applied Psychology*, 1966, **50**, 198–200.

Herzberg, F. The job attitudes in Soviet Union. *Personnel Psychology*, 1965, **18**, 245–252. (a)

Herzberg, F. The new industrial psychology. *Industrial and Labour Review*, 1965, **18**, 364–376. (b)

Herzberg, F., Mausner, B., & Snyderman, B. *The motivation to work.* New York: Wiley, 1959.

Kahn, R. L. Review of Herzberg, F., Mausner, B., and Snyderman, B. The motivation to work. *Contemporary Psychology*, 1961, **6**, 9–10.

Malinovsky, M. R., & Barry, J. R. Determinants of work attitudes. *Journal of Applied Psychology*, 1965, **49**, 446–451.

Maslow, A. H. *Motivation and personality.* New York: Harper, 1964.

Myers, M. C. Who are your motivated workers? *Harvard Business Review*, 1964, January–February, 73–88.

Porter, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *Journal of Applied Psychology*, 1962, **46**, 375–384.

Porter, L. W. Job attitudes in management: II. Perceived importance of needs as a function of job level. *Journal of Applied Psychology*, 1963, **47**, 141–148.

Porter, L. W., & Lawler, E. E., III. Properties of organizational structure in relation to job attitude and job behavior. *Psychological Bulletin*, 1965, **64**, 23–51.

Saleh, S. D. A study of attitude change in the preretirement period. *Journal of Applied Psychology*, 1964, **48**, 310–312.

Schwartz, M. N., Jenusaitis, E., & Stark, H. Motivational factors among supervisors in the utility industry. *Personnel Psychology*, 1963, **16**, 45–53.

Sinha, D. Job satisfaction in office and manual workers. *Indian Journal of Social Work*, 1958, **19**, 39–46.

Vroom, V. H. *Work and motivation.* New York: Wiley, 1964.

Vroom, V. H., & Maier, N. R. F. Industrial social psychology. In, *Annual Review of Psychology*, 1961. Palo Alto: Annual Reviews, Inc., 1961. Pp. 413–446.

Wernimont, P. F. Intrinsic and extrinsic factors in job satisfaction. *Journal of Applied Psychology*, 1966, **50**, 41–50.

(Received June 24, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 266-273

SVIB KEY LENGTH:

DISSIDENT DATA

NORMAN M. ABRAHAMS [1]

United States Naval Personnel Research Activity, San Diego

Recent studies of interest key length suggest that about 60 item responses are optimal for maximizing group preparation. This study sought to determine whether this optimum still emerges when control of item validity is accomplished by randomly selecting the item responses to be included in keys of various lengths. The possibility of an optimal key length was evaluated in terms of validity and test-retest reliability. These comparisons indicated that validities, as well as test-retest reliabilities, substantially and consistently increased as more item responses were scored. It was concluded that no optimal key length, such as the one previously proposed, exists when various item characteristics are randomized; consequently, key length is a decision specific to each item pool.

Over two decades ago, without the benefit of IBM, E. K. Strong, Jr. took the first empirical stand on how many item responses an occupational key should contain. On the basis of studies reported in his 1943 book, he emphasized the necessity of using long keys. Specifically, he suggested that every item response showing a percentage difference of six or greater, between the criterion group and men in general, be scored, using a variable weighting scheme. This procedure often resulted in scoring 200 or more item responses for a single key.

More recent studies (Clark, 1961; Strong, Campbell, Berdie, & Clark, 1964) indicate that it is unnecessary to score such a large number of item responses. Their comparisons show that shorter unit-weighted keys are as effective as longer ones in terms of both validity and reliability. In fact, they suggest that there is a maximum number of item responses to include in a scoring key, beyond which validity is reduced and test-retest reliability does not improve. With both Minnesota Vocational Interest Inventory (MVII) and Strong Vocational Interest Blank (SVIB) data this upper limit appears to be around 60–80 item responses.

A limitation in these recent studies is that test-retest reliabilities of short and long keys are based only on *short* time intervals. The crucial reliability question in using interest keys for long-range vocational guidance is: Are the short keys as reliable as long keys over *extended* periods of time? Thirty-day retest reliabilities of short and long keys have shown little differences. As yet, however, no studies have provided any assurance that short keys are as reliable as long keys over several years.

A further limitation in these studies is that the key-length comparisons confound the effects of "length of key" with the effects of "kind of key." Keys differ in "kind" in that they have different cutoff points—based on percentage differences—for item inclusion. And, they differ in "length" only indirectly as a function of these cutoff points; that is, the more stringent the percent difference cutoff point is, the fewer item responses available for scoring. As a consequence the shorter keys contain these plus an increasing number of progressively less-valid item responses. From this procedure it is impossible to determine whether the results are due to an *optimal* key length or, as Berdie and Campbell (1966) have suggested, to the fact that the keys in the so-called optimal range contain only the most-*valid* item responses.

The question that remains, then, is whether the reported superiority of the 60–80 item-

[1] Based on a paper presented to the University of Minnesota Center for Interest Measurement Research Symposium, May 1966. The opinions expressed are those of the author and do not necessarily reflect those of the Navy Department

response keys is due to key length per se, or to the fact that these keys contain the more-valid item responses. If key length is the reason, then the optimal range of 60–80 item responses may be used as a general guideline in constructing occupational keys. However, if the reason is the number of good items available, then optimal key length would be specific to each key-construction situation and will merely reflect how many good item responses can be found. Longer keys would then be constructed where the pools are rich in valid items and short keys where few good items are available.

Nash (1964), aware of the importance of an answer to this question, attempted to establish which factor was responsible for the optimal key length. He compared the validities of various length keys comprised of items equivalent in validity. Unfortunately, however, he used number of items, rather than item responses, as an index of key length. Since it is possible to score from one to three responses for each item, for the typical scoring key it would, in fact, be very unusual for these indexes to be the same. This makes Nash's data difficult to compare with the previous studies.

Given that a key is developed by first ascertaining which item responses are differentially endorsed by the criterion and reference groups—how many of these should be included in such a key? At one extreme there should not be so few as to produce an unreliable measure, and, at the other, not so many as to emphasize chance differences. While a wide range may exist between these extremes, the decision made here determines to a great extent the key's effectiveness. Even so, the evidence bearing on this important decision is conflicting and inconclusive.

The questions raised by the recent key-length studies provided the impetus and objectives for this investigation. Specifically, the questions are: Is the superior validity of the shorter keys a result of the *number* of item responses in a key, or is it a result of shorter keys having the better items? Is there a maximum number of scorable item responses, beyond which key validity is actually reduced, or are the findings due to the fact that only the best items have been used in the keys containing 60–80 items? And, last: Are the long-range retest reliabilities of short keys as high as those of long keys?

To answer these questions it was necessary to use a procedure that would allow comparing the validities of keys differing in length but similar in all other respects. Briefly, this

TABLE 1

Characteristics of the Three Criterion Groups Investigated

Characteristics	Group: Psychologists	Engineers	Osteopaths
Age	44	43.9	37.8
Average education	PhD	15.4	15.8
Minimum years in occupation	3	3	3
Sample size	1045	513	585
Development	697	342	390
Cross-validation	348	171	195

TABLE 2

Number of Experimental Keys Developed for Each Key Length within Each Psychologist Item Pool

Key length	Pool: 10–20	15–25	20–30	25–35	30–40
5	25	25	25	25	5
10	25	25	25	25	
15	25	25	25	25	5
20	25	25	25	25	
24					1
25	25	25	14[a]	25	
30	25	25	25	25	
35	25	25	25	25	
40	25	25	25	8	
50	25	25	25	2	
59				1	
60		2	2		
80	2	2	2		
93			1		
100	2	2			
150	2	2			
162		1			
200	2				
250	2				
255	1				

[a] Reduced number of keys at this length due to computer printing malfunction.

TABLE 3

NUMBER OF EXPERIMENTAL KEYS DEVELOPED FOR EACH KEY LENGTH WITHIN EACH ENGINEER ITEM POOL

Key length	Pool			
	10–20	15–25	20–30	25–35
5	25	25	25	25
10	25	25	25	25
15	25	25	25	11
16				1
20	25	25	25	
25	25	25	25	
30	25	25	5	
31			1	
35	25	25		
40	25	25		
50	10	10		
60	5	5		
80	5	5		
100	5	5		
103		1		
150	5			
200	5			
250	5			
256	1			

was accomplished by a random item-selection procedure that provided for different length keys equivalent in item validity.

PROCEDURE

The SVIB responses of current criterion groups and of the men-in-general group furnished the data on which this investigation was based.[2] These groups, as listed in Table 1, consisted of 1,045 psychologists, 513 engineers, 585 osteopaths, and 500 men in general.

For each occupational group, one-third of the total sample served as the hold-out cross-validation sample and two-thirds of the sample remained for item analysis. Sample sizes are shown in Table 1. The reference group—men in general—contained the SVIB responses of 500 men from a wide variety of occupations and was used for comparisons with each development and cross-validation sample. A group of Stanford University students retested after a 20-yr. interval provided SVIB responses for computing test-retest reliabilities.

From each validation sample, the percent difference of each item response was determined by subtracting the percentage of the reference group that endorsed the response from the corresponding percentage in the validation sample. In the usual manner, these percent differences provided the necessary data for assigning unit weights. All item responses having positive percent differences, that is, those that the validation sample endorsed more frequently, received weights of +1. Likewise, item-responses having negative percent differences received −1 weights.

The next step required setting up various "item pools" for each occupational group. For each group, relatively homogeneous item pools were established by specifying various upper and lower limits of item validity—in this case percent difference—and including in the pool only those item responses having a percent difference within the specified range. For example, for each occupational group all the item responses showing percent differences—both positive and negative—of 10–20 were placed in the low validity pool. The next pool contained all item responses having percent differences between 15 and 25, and so on. The percent difference limits of the pools used, as shown in Tables 2, 3, and 4, were 10-20, 15–25, 20–30, 25–35, and 30–40. Not all criterion groups had enough item responses showing large differences for the pools requiring the higher percent differences. In those cases, key-length comparisons could not, of course, be made.

Item responses were randomly selected from each pool for each of several different key lengths, ranging from five up to the total number of item responses in the pool. Tables 2, 3, and 4 show for psychologists, engineers, and osteopaths, respectively, the number of keys constructed at each length within each pool. The longest key shown for each pool is, of course, equivalent to the number of item responses in that pool. The resulting keys were validated and cross-validated on the development and hold-out samples. Only those keys obtained from the same

TABLE 4

NUMBER OF EXPERIMENTAL KEYS DEVELOPED FOR EACH KEY LENGTH WITHIN EACH OSTEOPATH ITEM POOL

Key length	Pool		
	10–20	15–25	20–30
5	25	25	25
10	25	25	25
15	25	25	25
20	25	25	25
25	25	25	4
30	25	25	1
35	25	25	
40	25	25	
50	25	15	
60	25	15	
70	4	1	
80	2		
100	2		
150	2		
172	1		

[2] The author is indebted to David P. Campbell for making the data for this investigation available

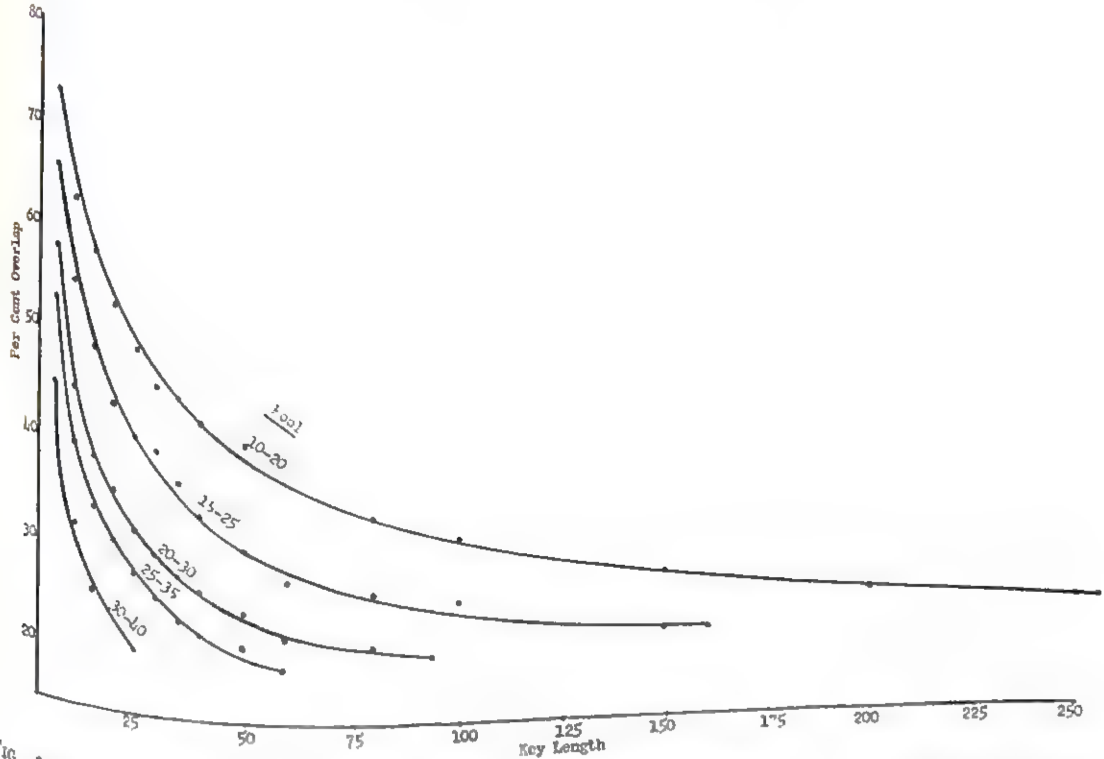


FIG. 1. Median percent overlap between reference-group distributions and psychologist validation sample distributions for the keys of each item pool as a function of key length.

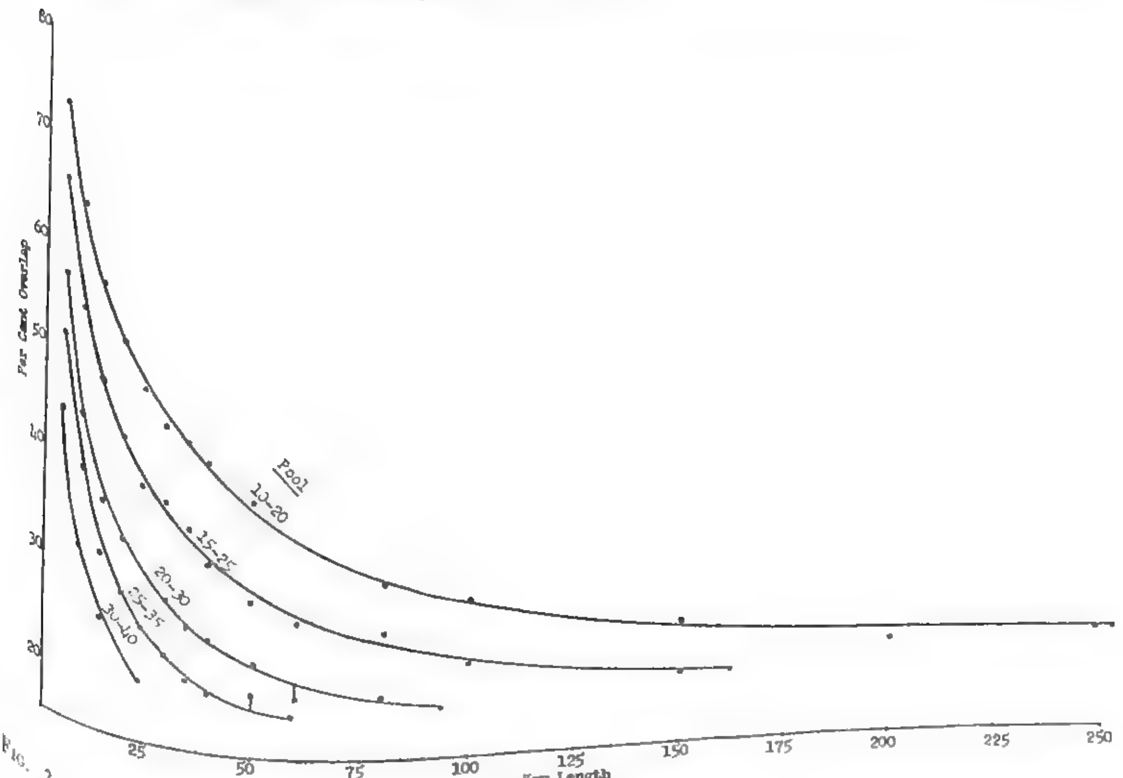


FIG. 2. Median percent overlap between reference-group distributions and psychologist cross-validation sample distributions for the keys of each item pool as a function of key length.

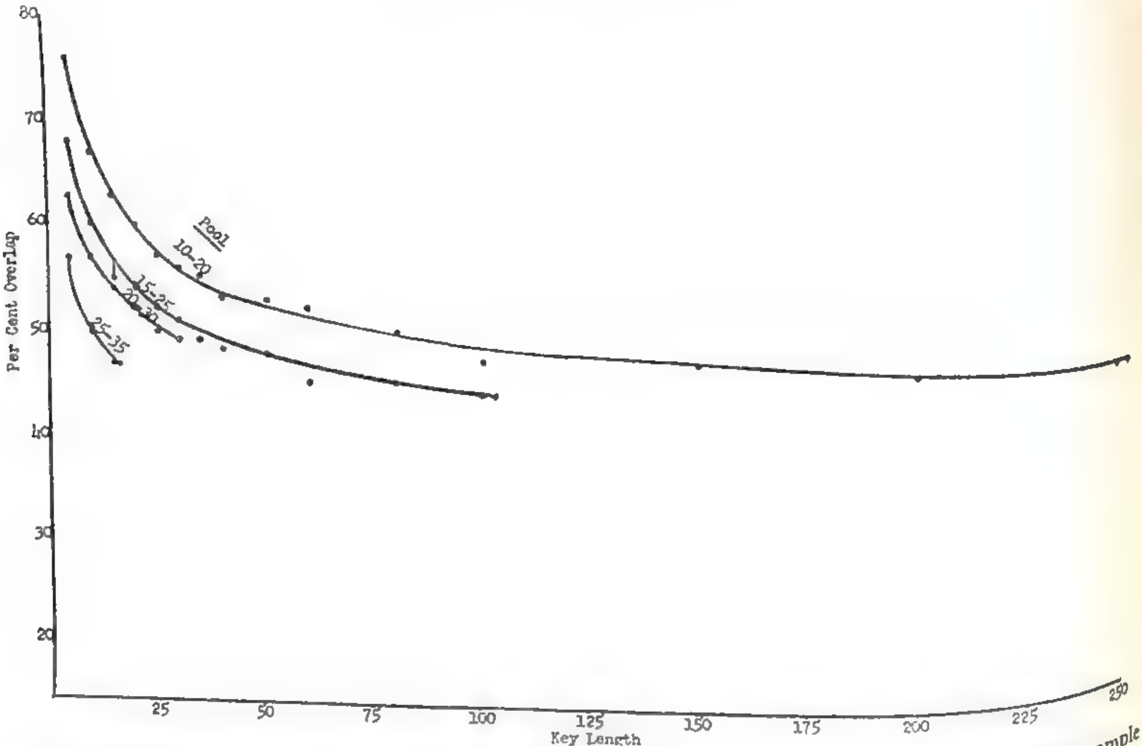


FIG. 3. Median percent overlap between reference-group distributions and engineer validation sample distributions for the keys of each item pool as a function of key length.

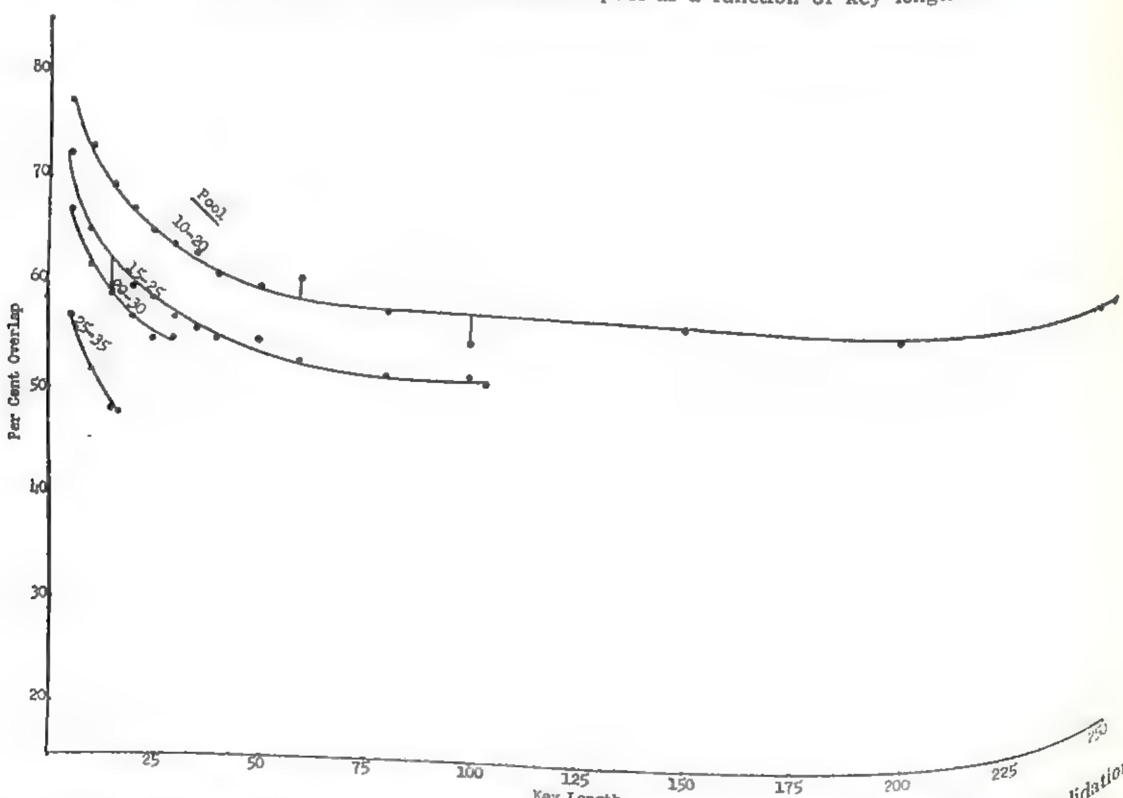


FIG. 4. Median percent overlap between reference-group distributions and engineer cross-validation sample distributions for the keys of each item pool as a function of key length.

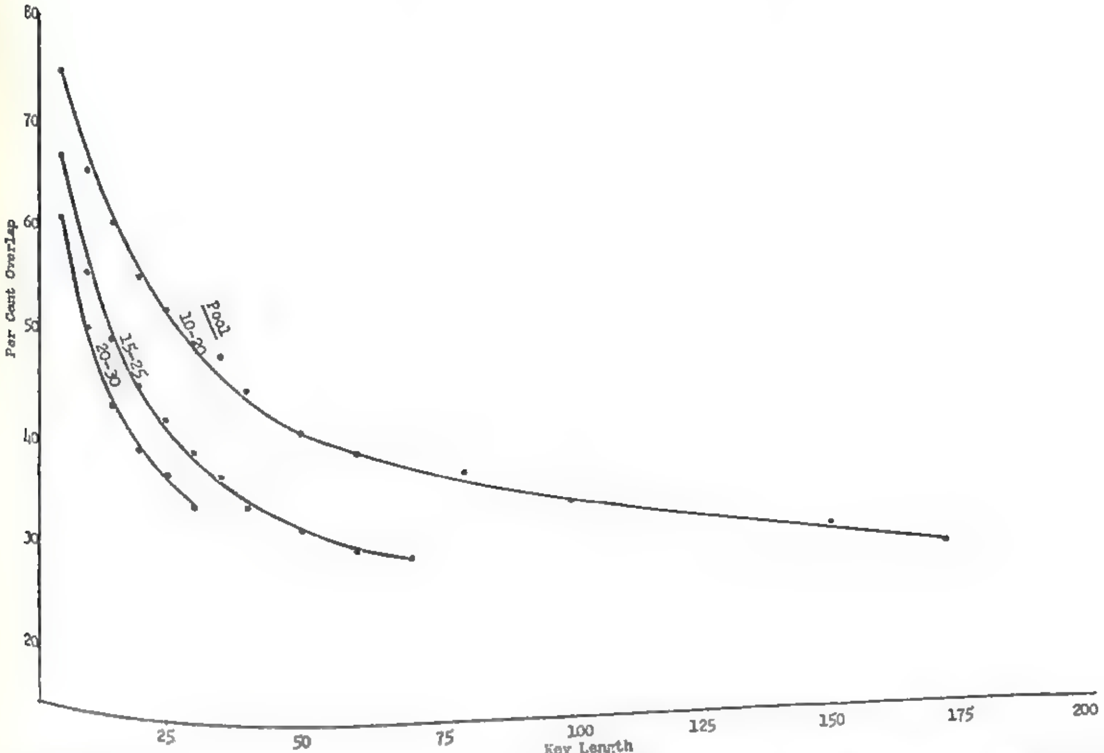


Fig. 5. Median percent overlap between reference-group distributions and osteopath validation sample distributions for the keys of each item pool as a function of key length.

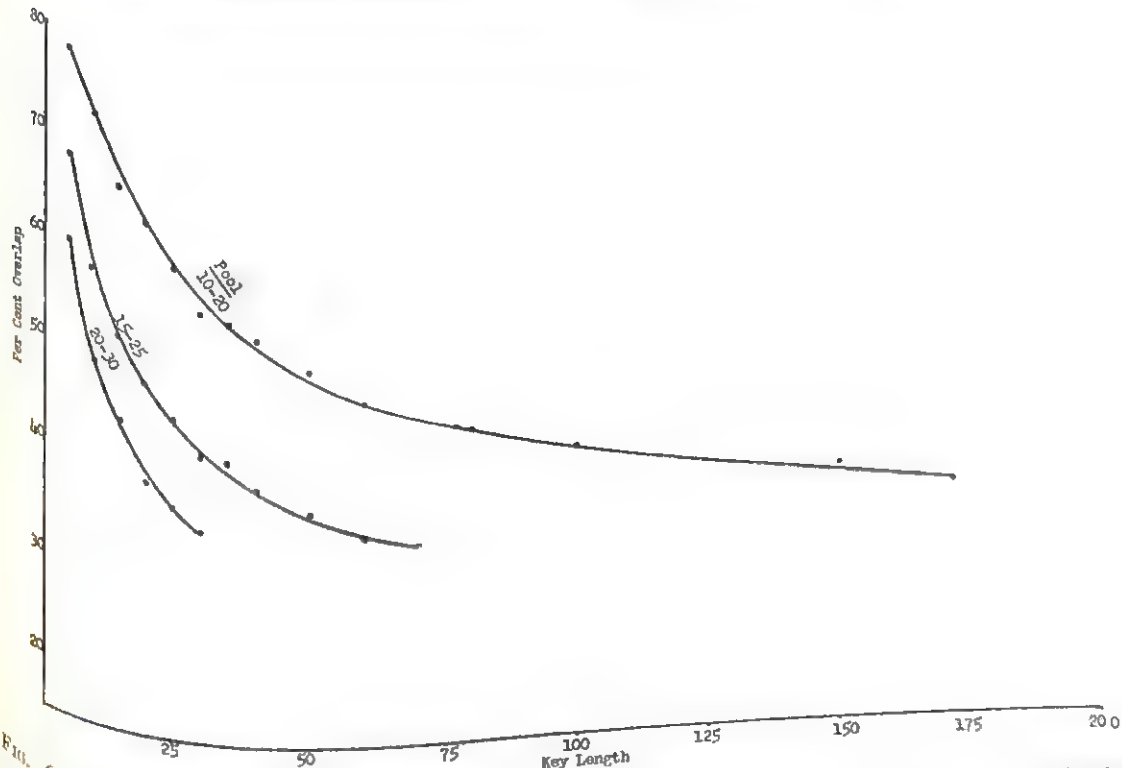


Fig. 6. Median percent overlap between reference-group distributions and osteopath cross-validation sample distributions for the keys of each item pool as a function of key length.

TABLE 5

RELIABILITIES AND TEST-RETEST MEANS AND STANDARD DEVIATIONS OF A 20-YR. RETEST GROUP (N = 197) FOR SELECTED KEY LENGTHS

Key length[a]	Test-retest correlation	Test M	Test SD	Retest M	Retest SD
25	.52	3.75	3.17	4.88	3.40
25	.52	3.06	3.11	4.82	3.18
30	.49	1.82	3.65	3.63	3.92
40	.50	3.41	4.22	5.43	4.88
50	.52	2.68	5.26	5.55	5.50
59	.57	3.40	6.07	6.97	6.65
25	.49	1.51	2.87	2.65	2.90
25	.53	−1.30	2.74	−.29	2.81
30	.47	2.02	2.86	3.08	3.39
30	.46	3.71	3.07	4.96	3.22
35	.50	−3.36	3.46	−1.99	3.57
35	.49	.51	3.61	1.52	3.66
40	.49	−.14	3.48	1.57	3.95
40	.53	−.32	3.71	1.05	3.73
50	.58	1.43	4.25	3.57	4.50
50	.60	2.69	4.38	5.04	4.36
60	.58	−.98	5.32	1.35	5.10
60	.51	−2.02	4.69	.99	5.41
80	.60	−3.22	6.34	−.18	6.82
80	.64	8.56	6.50	11.20	7.19
100	.60	−.29	7.74	3.82	8.76
100	.63	5.30	7.38	9.35	8.17
150	.69	1.74	10.42	9.06	11.84
162	.68	.75	11.20	7.81	12.61

[a] The first six keys were based on the psychologist 25–35% difference item-response pool, and the remainder were based on the 15–25 pool.

pool could be considered to have comparable item validity, thereby restricting key-length comparisons to keys based on the same item pool. Tilton's (1937) percent overlap, an estimate of the proportion of the scores in one distribution that match the scores in another distribution, was used as the measure of key validity. For the several keys at each length within each pool, a median validation and cross-validation percent overlap was computed. Next, for each item pool, median validity—in terms of Tilton's percent overlap—was plotted as a function of key length. Curves depicting these relationships for both validation and cross-validation groups are shown in Figures 1 through 6.

RESULTS AND DISCUSSION

Inspection of the curves in Figures 1, 3, and 5 reveals a monotonic decrease in percent overlap as the number of item responses that were scored was gradually increased from five up to the total number in the pool. As can be seen from Figures 2, 4, and 6, this relationship held up on cross-validation of the keys in every item-response pool and each criterion group. These data, then, clearly indicate the absence of an optimal upper limit on the number of item responses to score for obtaining maximum group separation when various item-response characteristics are randomized across different key lengths. In fact, these data indicate that considerable gains, rather than losses, are made by increasing key length beyond 60 item responses. These increases could not be ascribed to higher item validity or any other item characteristic since these effects were randomized across keys within each item pool. However, the curves show that as more item responses are scored, advances made in validity become increasingly smaller. This is probably due to the increasing difficulty of adding item responses that measure previously untapped variance.

These figures also show an apparent tendency for shrinkage upon cross-validation to increase as key length is increased. However, since the increments in key length consist of part-valid and part-error variance, increasing in unknown proportions, it is unclear whether or not the observed relationship between key length and shrinkage is an artifact.

To determine whether short keys are as stable as long keys over an extended period, reliabilities for a selected number of keys were computed for a group to whom the SVIB was readministered after a 20-yr. interval. From the large number of keys constructed for the psychologist criterion group, 24 were selected for reliability analysis. These keys ranged in size from 25 to 162 item responses and were taken from two different item-response pools. Of these, 18 were from the 15–25% difference pool and 6 were from the 25–35 pool. From the keys having a percent overlap representative of the median overlap at each length within these two pools, a random selection of one or two keys was made.

Only the keys within the 15–25 pool provided a wide enough range in length to make meaningful reliability comparisons. Although keys and their corresponding reliabilities for both item pools are shown in Table 5, the relevant ones, for present purposes, begin in the seventh row. Comparisons of these reli-

abilities indicate that the shorter keys account for about 25% of retest score variance over a 20-yr. interval, whereas the longest key accounted for about 50%. These data show a serious loss in 20-yr. retest reliabilities for shorter keys. Such results clearly indicate that if shorter keys are to be constructed, considerable attention should be devoted to their long-range stability.

Conclusions

With respect to the optimal-key-length question, the primary conclusions to be derived from these data are:

1. When item validity is controlled, as was done in this study, validity does not peak at any point but continues to rise as the key length increases, and no particular key length emerges as optimal.

2. With the larger item pools, a virtual asymptote appears at about 100–150 item responses.

3. The so-called "optimal point" of 60 item responses is *not* due to key length, but is probably due to the necessity of adding item responses of reduced validity and giving them weights equal to the "best" item responses. This result is, therefore, a consequence of unit weighting item responses and will occur in other situations although the actual length will vary as a function of item heterogeneity and validity.

4. When test-retest reliability is assessed over a 20-yr. interval, in contrast to the previous 30-day assessments, it appears that keys should contain at least 100 item responses.

References

Berdie, R. F., & Campbell, D. P. The measurement of interests. In D. K. Whitla (Ed.), *Handbook of psychological measurement.* Reading, Mass.: Addison Wesley, 1967, in press.

Clark, K. E. *Vocational interests of nonprofessional men.* Minneapolis: University of Minnesota Press, 1961.

Nash, A. N. Development and evaluation of a Strong Vocational Interest Blank key for differentiating between potentially effective and less effective business managers. Unpublished doctoral dissertation, University of Minnesota, 1964.

Strong, E. K., Jr. *Vocational interests of men and women.* Stanford: Stanford University Press, 1943.

Strong, E. K., Jr., Campbell, D. P., Berdie, R. F., & Clark, K. E. Proposed scoring changes for the SVIB. *Journal of Applied Psychology,* 1964, 48, 75–80.

Tilton, J. W. The measurement of overlapping. *Journal of Educational Psychology,* 1937, 28, 656–662.

(Received June 27, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 274-277

GOAL SETTING AS A MEANS OF INCREASING MOTIVATION[1]

JUDITH F. BRYAN AND EDWIN A. LOCKE

American Institutes for Research, Washington Office

On the basis of differences in performance in relation to maximal ability and differences in attitude ratings on an addition task, a low-motivation and a high-motivation group were selected for 2 retests on the same task. The low-motivation group was given specific goals to reach, and the high-motivation group was told to do their best on each trial of each retest. By the end of the 2nd retest, the group given specific goals had "caught" the Do-Best group both in terms of performance and in terms of favorable attitudes toward the task. The results suggested that specific goals can be used to motivate Ss who bring a low degree of motivation to the task situation.

Previous studies of the effects of goals on performance and attitudes have found: (*a*) that specific hard goals produce a higher performance level than a goal of "do your best" (Locke & Bryan, 1966a, 1966b, 1967; Mace, 1935); (*b*) that hard goals yield less overall task liking and satisfaction than easy goals (Locke, 1965, 1966b); and (*c*) that specific hard goals produce more interest (less boredom) in the task than "Do-Best" goals (Locke & Bryan, 1967). In all of the above studies, the methodology consisted of taking subjects (*S*s) "equated" on the dependent variables to begin with and attempting to produce differences in these variables as a function of the experimental conditions. The present study dealt with the same problems as those cited above, but the methodology chosen was exactly the opposite of the usual procedure. Instead of taking *S*s who were *initially similar* on the dependent measures and *making them different*, *S*s were chosen who were *initially different* on these variables and experimental procedures were directed toward *making them similar*.

The reason for this unusual procedure was to try to duplicate more closely (in one respect) situations typically found in industrial and military performance situations. Given the low validity of presently available tests of "motivation" (Guion & Gottier, 1965), it is difficult, if not impossible, to select workers on the basis of their "motivation" in advance; thus the employer or trainer is faced with the problem of motivating workers who may be quite heterogeneous with respect to the motivation they bring to the job situation. The specific problem then becomes one of raising the performance of the more poorly motivated individuals rather than one of selection.

Another difference between this and previous studies was the use of four different attitude measures. Except for one study by Locke (1966b), previous studies by the present authors had measured only one type of attitude (either liking or boredom-interest). In this study four different attitudes were measured: boredom-interest, degree and intensity of task focus (concentration), and experienced effort. It was predicted that all four types of attitudes and performance would be enhanced by giving *S*s specific goals as compared to a Do-Best goal.

PROCEDURE

Subject Pool

The original *S* pool consisted of 10 male and 10 female paid college student volunteers from the University of Maryland.

Task and Rating Scales

The task was simple addition, each problem consisting of three two-digit numbers. Each *S* was presented with a booklet containing a separate sheets(s) of addition problems for each trial. The number of problems per page varied according to trial length and to the arrangement of the problems on the page. The arrangement was such as to prevent *S*s from keeping track of their scores.

Separating each trial sheet was a page consisting of four attitude rating scales. The Interest-Boredom

[1] This research was supported by Contract Nonr 4792 (00) between the Office of Naval Research and the American Institutes for Research. The authors would like to thank William J. Baker of AIR for his many helpful comments and suggestions relating to the statistical analysis of the data.

scale was a bipolar, vertical, graphic scale anchored at various points by statements ranging from "it was extremely boring" to "it was fascinating." There were two Task Focus scales, one asking for the *percentage of time* S felt he was focused on the task and the other asking how *intensely* S felt he was focusing. The Effort scale asked S to indicate the percentage of his maximum possible effort exerted during the trial. The S was told to rate the experience of effort and not to make an induction from performance.

There were 12 trials of steadily increasing length. Trials 1–4 lasted 15 sec., Trials 5–6 were 30 sec. long, followed by Trials 7–12 which were, respectively, 1, 2, 4, 8, 16, and 32 min. in length. The Ss had 30 sec. between trials to fill out the four rating scales. The purpose of the differing trial lengths was to allow the selection of Ss who were similar in performance rate on the shorter trials ("maximal ability") but different on the longer trials, a drop on the latter being taken as indicative of lowered motivation to perform.

Method

Test I. The experiment was introduced as a study of attitudes as a function of trial length. All Ss were instructed to "do their best." It was added that they should try to work at a steady pace, exerting the same amount of effort per time unit on every trial.

Of the original 20 Ss taking Test I, 12 were selected for Tests II and III. The six Ss having the uniformly lowest performance slopes and lowest scores on the attitude scales were placed in the *Goal* condition for the next two tests. The six Ss having the uniformly highest scores on these measures were placed in the *Do-Best* condition for Tests II and III. The differences between the two groups on total scores on the rating scales on Test I were all significant at the .05 level or better (t's $= 2.49$, 3.15, 3.63, and 4.20 for the boredom-interest, percent focus, intensity of focus, and effort measures, respectively). The difference between the groups in total performance on Test I was not significant ($t = 1.34$), but generally the performance rate (in terms of mean problems correct per minute) of the low-motivation Ss decreased more than did the performance of the high-motivation Ss (the two groups had identical performance means on the four 15-sec. trials). Thus the two groups were of the same "maximal ability" as indicated by their performance on the shortest trials, but the one group dropped off more on the longer trials and showed less interest in and focus on the task on Test I, as a whole, than did the other. The general design of the experiment is shown in Table 1.

Test II. Two to three wk. after taking Test I, the Goal and Do-Best Ss returned for Test II, which was identical to Test I. The instructions for the Do-Best group were the same as those given to all Ss before Test I, though it was suggested that they might find the going a little faster this time since they had worked these same problems previously. The Goal Ss were told that on Test I they had slowed down in performance during the longer trials as compared to the Do-Best Ss and to try and improve their performance on Test II by trying to reach the specific goals which had been set for them on each trial. They were told that the goals on the short (15-sec.) trials were set at about the same place that they had reached on Test I, but that the goals for the longer trials were set above what they had done previously, since their performance had been poorer on these trials.

The goals for the Goal group were marked by a red circle placed around the problem the S was to try to reach by the end of the trial. These goals were determined by taking the mean number done correctly by S on the four 15-sec. trials on Test I as the base rate and making each subsequent goal an appropriate multiple of this so that S would have to maintain his initial rate to reach each subsequent goal.

Test III. Two to five wk. ($M = 2.5$) after Test II, the Goal and Do-Best Ss returned to take the identical test again. The instructions for the Do-Best group were the same as those used in Tests I and II. The Goal Ss were told that they had raised their performance on Test II but to try to further improve their performance on Test III by trying for new goals. These new goals were set 10% above their own best previous performance (on Test I or II) on each trial. (The goals were again marked by red circles on the worksheets.) This new method of setting goals was used here, because it was found that the method used for Test II yielded goals that were too hard on the longer trials.

TABLE 1

GENERAL DESIGN OF THE EXPERIMENT

Group	Test I	Test II	Test III
Low motivation[a]	Do Best	Goal	Goal
High motivation[a]	Do Best	Do Best	Do Best

[a] Motivation groups established on the basis of performance slopes and attitude ratings on Test I.

RESULTS

The Ss in the Goal group were able to reach or beat their goals 46% of the time on Test II and 62% of the time on Test III. The scores used in all analyses (with two exceptions to be described below) were total problems attempted, total problems correct, and the total interest, focus, and effort ratings for all trials combined for each test.

TABLE 2

*F*s FOR AB INTERACTION BETWEEN GROUPS ACROSS TESTS AND FOR TEST EFFECTS WITHIN GROUPS

Measure	F for AB interaction[a]	F for test effects within groups[a]	
		Do Best	Goal
Problems attempted	3.92*	0.88	7.93**
Mean of mean attempted	10.00***	4.62*	24.94***
Problems correct	2.79	1.55	8.18**
Mean of mean correct	6.20**	8.58**	23.18***
Interest	2.45	3.03	1.07
Percent focus	4.56*	3.06	1.82
Intensity of focus	6.69**	6.27**	1.43
Effort	5.72*	3.02	2.89

[a] df for all Fs 2/20.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table A[2] shows the mean correlations (Pearson *r*'s and *rho*'s) of these total scores across tests for each group. The Goal group shows relatively high consistency on all the measures across tests, while the Do-Best group shows consistency only on the two performance measures. (There is no obvious explanation for their low consistency on the attitude measures.)

Table B (see Footnote 2) is a summary matrix showing the mean Pearson correlations among the various performance and attitude measures. For each group, the correlations among the raw total scores were computed separately for Tests I, II, and III; and the correlations among change scores were computed using the Test III–I differences. The summary matrix is the average of these eight 6 × 6 matrices and, with one exception, is representative of the individual matrices. There was a high (though spurious) correlation between problems attempted and problems correct. There was no relation between either of the performance measures and any of the attitude measures, suggesting that performance and attitudes represented separate aspects of motivation. On the other hand, three of the four attitude measures are highly related to each other, interest-boredom being the exception.

Mean total scores for each group on each test for each of the six measures are shown in Figures 1, 2, and 3. Generally, all curves show convergence across the three tests, with the Do-Best group decreasing in performance, interest, focus, and effort, and the Goal group increasing on these measures. There were no significant differences between the groups in

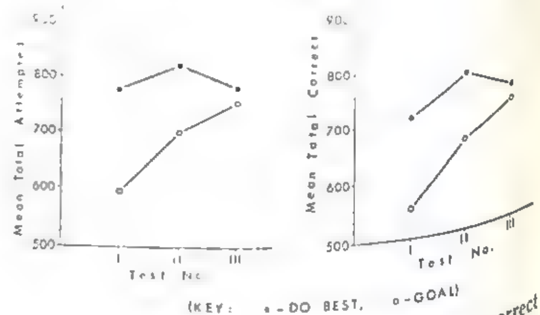


FIG. 1. Mean total attempted and mean total correct of do-best and goal groups by test.

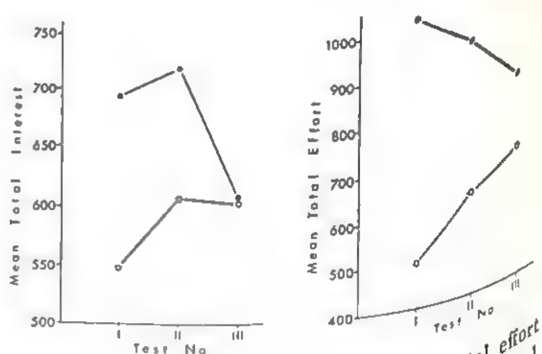


FIG. 2. Mean total interest and mean total effort of do-best and goal groups by test. (See Figure 1 for key.)

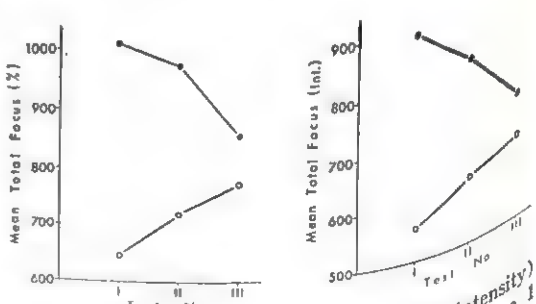


Fig. 3. Mean total focus (percentage and intensity) of do-best and goal groups by test. (See Figure 1 for key.)

[2] Tables A and B have been deposited with the American Documentation Institute. Order Document No. 9391 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $1.25 for microfilm or $1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

mean scores on any measure on Test III, whereas all corresponding differences except performance were significant on Test I. On the last one or two trials of Test III, the Goal group's mean actually exceeded the Do-Best group's mean on interest, intensity of focus, and problems correct.

To further analyze these changes, a two-factor repeated-measures design (Winer, 1962, p. 302ff) was used. The *F* ratio of interest in the present case was the AB interaction which reveals any differences in trends of the Goal and Do-Best groups across tests. In addition to using total scores on the performance measures, the mean of the individual trial means for each of these measures was used, the purpose being to weight all trials equally as was the case with the attitude measures. The *F* ratios for the eight resulting measures are shown in Table 2. Also shown are the *F*s for changes across tests for each group.[3] All *F*s for interaction, except those for total correct and interest, were significant at the .05 level or better. (The total correct and interest measures yield significant AB interactions at the .05 level when the above analysis was performed using the data for Tests I and III only.) The significant performance interaction was the result of a small quadratic trend in the Do-Best group versus a strong linear trend in the Goal group.[4] The attitude interactions resulted from linear trends in opposite directions for the Do-Best and Goal groups (see Footnote 4).

Discussion

The results of this experiment indicate that specific performance goals can serve to raise the motivation of *S*s who are low in the motivation they bring to the task. The assigning of specific and reasonably hard goals to these *S*s raised performance level and favored the development of more positive attitudes toward the task. On the other hand, telling the high-motivation *S*s to do their best resulted in little performance increase and the development of increasingly less favorable attitudes toward the task.

The lack of correlation between the performance and attitude measures supports the findings of previous research that attitudes and performance are not necessarily related (Brayfield & Crockett, 1955; Locke, 1966a; Locke & Bryan, 1967).

[3] The error term for these analyses was the same as for the AB interaction (B × Subjects within groups).

[4] The breakdown of the between-cell variance (across tests) showed virtually all the between-cell variance to be linear except for performance for the Do-Best group, which was predominantly quadratic.

REFERENCES

BRAYFIELD, A. H., & CROCKETT, W. H. Employee attitudes and employee performance. *Psychological Bulletin*, 1955, **52**, 396–424.

GUION, R. M., & GOTTIER, R. F. Validity of personality measures in personnel selection. *Personnel Psychology*, 1965, **18**, 135–164.

LOCKE, E. A. The relationship of task success to task liking and satisfaction. *Journal of Applied Psychology*, 1965, **49**, 379–385.

LOCKE, E. A. A preliminary study of some cognitive factors in boredom. American Institutes for Research, Silver Spring, Maryland, AIR-R&D31-3/66-TR, March 1966. (a)

LOCKE, E. A. The relationship of task success to task liking: A replication. *Psychological Reports*, 1966, **18**, 522–554. (b)

LOCKE, E. A., & BRYAN, J. F. Cognitive aspects of psychomotor performance: The effects of performance goals on level of performance. *Journal of Applied Psychology*, 1966, **50**, 286–291. (a)

LOCKE, E. A., & BRYAN, J. F. The effects of goal-setting, rule-learning, and knowledge of score on performance. *American Journal of Psychology*, 1966, **79**, 451–457. (b)

LOCKE, E. A., & BRYAN, J. F. Performance goals as determinants of level of performance and boredom. *Journal of Applied Psychology*, 1967, **71**, 120–130.

MACE, C. A. Incentives: Some experimental studies. Industrial Health Research Board (Great Britain), 1935, Report No. 72.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill 1962.

(Received June 29, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 278–283

CHARACTERISTIC PACE AS DETERMINED BY THE USE OF A TRACKING TREADMILL

GARY L. HOLMGREN[1] AND GEORGE S. HARKER

United States Army Medical Research Laboratory, Fort Knox, Kentucky

It was found that when Ss walked on a tracking treadmill under a "comfortable-but-determined" (C-D) walking instructional set for a minimum of 30 min. on each of 3 testing days: (a) Ss demonstrated a characteristic C-D pace that was stable on any given day; (b) Ss' C-D pace differed statistically from each other ($p < .001$); and (c) Ss' C-D pace measures were most reliable between Testing Days 2 and 3 ($p < .01$).

A growing interest in and understanding of various modes of muscle activity has recently been experienced. The study of the characteristics of limb activity has been expanded to include: (*a*) the endurance of the leg as a function of the force applied to a pedal (Elbel, 1949); (*b*) the relationship between initial maximum grip strength and strength endurance for a fixed period of time (Tuttle, 1950); (*c*) the response endurance of different muscle groups at various fractions of the maximum response strength (Rohmert, 1960); (*d*) the temporal components of motion in human gait (Smith, McDermid, & Shideman, 1960); (*e*) the multidimensional force components of "normal" and hemiplegic gait (Barany, 1964); (*f*) the force components of deep knee bends (Ismail, 1964); and (*g*) the proportionate relationship of the sustained grip force to the maximum strength of the response as an index of working efficiency and recovery function for various relative loads (Caldwell, 1964).

There have been two exciting innovations in the above-mentioned research: The first being Barany's (1964; Barany & Greene, 1961) construction of a force platform suitable for measuring the bodily forces exerted by hemiplegic patients while walking on a treadmill and the subsequent development of a data reduction system which accurately integrates the resulting force traces over time with respect to amplitude and direction (frontal, lateral, and vertical). The second innovation is found in Caldwell's (1963, 1964) research in which he has refined a technique employing relative loading which resulted in the demonstration that differences in grip endurance were unrelated to differences in strength.

The prior treadmill research conducted in this laboratory by Evans (1962, 1963), who developed a technique utilizing a titration schedule to measure performance decrement as a function of continuing heavy muscular exertion, and Caldwell's research, utilizing a static manual response, led to the idea that the treadmill could conceivably be utilized to investigate the application of the "relative or proportional loading" technique to the dynamic muscle group responses involved in walking.

Before relative loading can be applied to pace, a valid and reliable point of reference must initially be determined (e.g., analogous to Caldwell's maximum grip strength). Initially it was felt that the point at which the subject (*S*) broke into a run would yield such a measure. Several pilot studies, however, suggested that such a measure may be lacking in the necessary validity and reliability for a variety of reasons: (*a*) the effect of the treadmill task just preceding the maximum walk measure was a confounding factor; and (*b*) the particular instructions given *Ss* were found to influence the maximum walking rate, and, among other problems, the treadmill maximum speed of 6.5 mph did not exceed the speed of the fastest walker.

In view of the above it was determined to evaluate *Ss*' "comfortable-but-determined" (C-D) walk in the hope that a valid and reliable "characteristic" pace might be obtained. If it could be demonstrated that *Ss* have a stable characteristic pace then it would be

[1] Now psychology research director at Texas Instruments, Inc., Dallas, Texas.

possible to increase or decrease the pace by a percentage to study various measures of strength and endurance.

The purposes therefore of the present research were to determine: (*a*) if individuals have a characteristic pace; and (*b*) if the pace is stable over time.

METHODS AND PROCEDURES

Apparatus

The apparatus was essentially that described by Evans (1963), with a few modifications. The variable speed (1.5–6.5 mph) treadmill was set at a 0.0% grade. The speed control of the treadmill, which in the usual application is manually set to a fixed speed for a run, was connected by an appropriate mechanical coupling to a reversible electric motor. In one direction of rotation this motor drove the speed control of the treadmill to produce a constant acceleration of the treadmill. In the other direction of rotation the speed control produced a constant deceleration. The speed-control motor was set to produce an acceleration or deceleration of the treadmill of 2 ft/min/sec.

In the Evans (1963) study, the direction of rotation of the speed-control drive was determined by a spring-loaded switch held by *S* as he walked on the treadmill. When *S* depressed the switch, the treadmill decelerated; when *S* released the switch, the treadmill accelerated. This arrangement required that *S* continuously make overt responses to the acceleration or deceleration of the treadmill to produce the titration schedule. Under these conditions an *S* instructed to walk at a C-D pace would establish his "steady" pace by constantly accelerating or decelerating the treadmill to produce a sinusoidal envelope of treadmill speeds.

In the present study it was desired that the treadmill speed "track" *S*'s pace as he walked on the treadmill. This was achieved by the use of a three-position spring-loaded switch actuated from a belt clip attached to *S*. The system was so designated that if *S* walked at a designated location on the treadmill the switch would be in the neutral (Number 2) position, and the acceleration-deceleration (A-D) drive would not be activated, so the treadmill would run at a constant speed. If, however, *S* increased his pace (either length of stride or steps per minute), since he would move forward on the mill, the switch would be pulled to the accelerate (Number 1) position, and the A-D drive would accelerate the treadmill to match speed with *S*, at which time the walker would again be at the neutral location on the treadmill allowing the switch to drop into the constant velocity position. Conversely, if *S* decreased his pace he would simultaneously move backward on the treadmill, the switch would drop to the decelerate (Number 3) position, and the A-D drive would decelerate the treadmill until its speed matched that of *S*'s pace. Earphones with muffs and a low-level (10 db.) white noise were employed to mask the noise of the activated A-D drive to insure that the primary cues to walking pace would be of a proprioceptive mode. To further control extraneous cues, *S* was instructed to look straight ahead and the manual speed change handle on the side of the treadmill was covered, as was the front edge of the treadmill belt.

A continuous tracing of treadmill speed obtained from a tachometer generator attached to the treadmill drive was recorded graphically with a Moseley Model 680 strip chart recorder.

The advantages of the above procedures are: (*a*) *S* is not forced to respond to the continual A-D implicit in the titration schedule approach; (*b*) the speed track of the treadmill should be narrower, more closely following the natural accelerations and decelerations of normal walking; and (*c*) extraneous cues which might influence the walking pace are minimized.

Subjects

Nineteen male *Ss* who had just completed 6-mo. basic training were employed in this investigation. The *Ss*, whose average age was 19.7 yr., wore regulation fatigue pants, tee shirts, and combat boots (average weight 5.5 lbs.). Walking was performed in a temperature-controlled chamber at 80° F.

Procedure

During the week prior to the data-taking sessions, each *S* walked for 30 min. on the treadmill for familiarization. The purpose of the familiarization was to allow *Ss* to become adapted to walking on the treadmill and to become adjusted to the accelerating and decelerating of the treadmill as it matched their pace.

The task for all *Ss* on each of the 3 days of testing was as follows: (*a*) two short walks starting from either a fast walk (5 mph) or a slow walk (2 mph) during which *Ss* either decelerated or accelerated their pace to reach their "comfortable-but-determined" (C-D) walking pace; (*b*) a 30-min. walk at their C-D pace; and (*c*) two short walks again starting from either a fast or a slow walk to adjust their pace to reach their C-D pace. For Group I, the short pre- and postwalk (fast start then slow start or slow start then fast start) was randomly determined and the walk lasted 4 min. For Group II, the pre- and postwalking conditions were randomized but the walk lasted 8 min. The present concern is with only the 30-min. walk.

The instructional set for the C-D pace was determined by the instructions "walk as though you were going from here to the theater (.5 mi.) and had just time enough to get there before the start of the feature without pushing yourself too hard." There was a 5-min. rest between each condition. The session for each of the 3 testing days was 66 min. for *Ss* in Group I and 86 min. for *Ss* in Group II.

TABLE 1

SUMMARY TABLE OF ANALYSIS OF VARIANCE OF THE FIRST 15 MIN. VERSUS THE LAST 15 MIN. OF THE C-D WALK

Source	*df*	*MS*	*F*
Group I			
Days (A)	2	2.6644	1.3989
Conditions (B)	1	.1285	.8718
Subjects (C)	8	15.7738	1421.0631*
A × B	2	.0446	.4811
A × C	16	1.9047	171.5946*
B × C	8	.1474	13.2793*
A × B × C	16	.0927	8.3514*
Within	756	.0111	
Total	809		
Group II			
Days (A)	2	1.8976	.4022
Conditions (B)	1	.0205	.0414
Subjects (C)	9	21.4918	820.2977
A × B	2	.4124	1.5451
A × C	18	4.7176	180.0611*
B × C	9	.4954	18.9084*
A × B × C	18	.2669	10.1870*
Within	840	.0262	
Total	899		

* $p < .001$.

RESULTS AND DISCUSSION

In an effort to learn if *Ss* revealed a characteristic pace during their 30-min. walk and if this pace was stable over time, two analyses of variance were computed (Table 1).

The analyses reveal that the main-effect conditions (first 15 min. versus last 15 min. of the 30-min. walk) were not significant, suggesting that for each group the C-D pace an *S* adopted was relatively stable. The *Ss'* main effect was found to be highly significant ($p <$ .001) as was the *Ss* × Days interaction. These findings reveal that *Ss'* paces differ from each other and suggest that though an *S's* pace was stable for any one day, his pace may differ from day to day. These observations contribute to accounting for the fact that both the Conditions × *Ss* and the Days × Conditions × *Ss* interactions were statistically significant. No systematic day to day variation in *Ss'* paces (such as a continual increase or decrease in pace) was observed as evidenced by the fact that neither the main-effect days nor

TABLE 2

MEANS AND STANDARD DEVIATIONS FOR EACH SUBJECT'S 30-MIN. C-D WALK FOR DAYS 1, 2, 3, AND AVERAGE OF DAYS 1–3 IN MPH

Subject	*M*				*SD*			
	Day 1	Day 2	Day 3	Days 1–3	Day 1	Day 2	Day 3	Days 1–3
Group I								
1	2.72	2.40	2.15	2.42	.142	.118	.152	.270
2	3.17	3.17	3.00	3.11	.122	.113	.069	.133
3	3.04	2.65	2.95	2.88	.063	.128	.078	.194
4	3.03	3.83	3.46	3.44	.082	.136	.103	.343
5	3.16	3.34	3.47	3.32	.089	.086	.118	.160
6	2.96	2.15	2.33	2.48	.120	.178	.091	.375
7	3.34	3.00	2.88	3.07	.113	.116	.138	.231
8	3.52	3.35	3.20	3.36	.079	.111	.237	.204
9	3.76	3.62	3.53	3.63	.070	.075	.114	.131
$\bar{X}$	3.19	3.05	3.00	3.08	.098	.118	.122	.227
Group II								
1	3.94	4.02	3.87	3.94	.202	.045	.135	.157
2	3.77	3.78	3.98	3.85	.069	.114	.135	.145
3	3.61	2.83	2.84	3.09	.414	.184	.176	.452
4	3.88	3.79	3.91	3.86	.174	.083	.219	.177
5	3.79	3.74	3.88	3.81	.335	.110	.098	.218
6	2.83	2.97	3.23	3.01	.173	.135	.138	.232
7	3.79	3.87	3.79	3.82	.185	.100	.230	.183
8	2.56	2.23	3.06	2.62	.270	.275	.115	.413
9	2.95	4.17	4.54	3.89	.241	.107	.173	.701
10	3.75	2.80	2.68	3.08	.284	.124	.138	.518
$\bar{X}$	3.49	3.42	3.58	3.50	.236	.128	.156	.319

the Days × Conditions interaction were statistically significant.

In an effort to determine what relationships did exist among the 3 days of testing, a series of Pearson product-moment correlations were computed.

The average mph for the 30-min. walk was computed for each *S*. Means and standard deviations for *S*s in both groups are presented in Table 2. Justification for averaging across the 30 min. is taken from the observation that the Conditions main effect was not statistically significant. The mean mph for an *S* is therefore a measure of his C-D pace and was used to compute the correlations between days (Table 2).

Inspection of Table 3 reveals that for both groups the C-D pace was more reliable between Days 2 and 3 than between any other 2-day combination ($p < .01$). The analyses up to this point indicate that performance in each of the two groups was comparable with the exception that the average mph for Group I was .42 mph lower than Group II. An even more interesting difference, however, is revealed by the fact that the correlation between Days 1 and 3 for Group I was significant beyond the .05 level of significance, but for Group II the comparable correlation did not even approach significance (Table 3).

Inspection of Figure 1 reveals that on Day 1 *S*s' performance was more variable as compared to Days 2 and 3 for Group II but not for Group I, which accounts for the lower Day 1 versus Day 2 and Day 1 versus Day 3

TABLE 3

CORRELATION MATRIX OF SUBJECTS' C-D PACE ($\bar{X}$ MPH FOR 30 MIN.) FOR THE 3 TESTING DAYS

Group	Day 2	Day 3
I		
Day 1	.61	.67*[a]
Day 2		.92**[a]
II		
Day 1	.51	.14
Day 2		.90**[b]

[a] $r > .798$.
[b] $r > .765$.
* $p < .05$.
** $p < .01$.

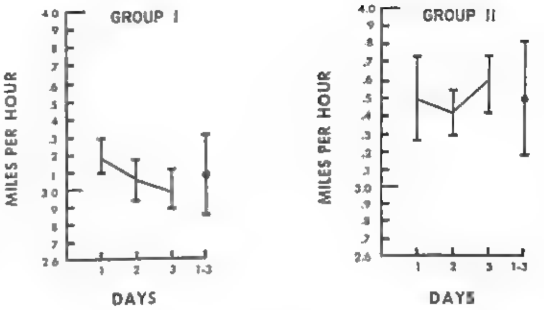


FIG. 1. Mean and standard deviation of 30-min. walk for the 3 testing days. (Groups I and II)

correlations for Group II as compared to Group I.

Because the correlations between Days 2 and 3 were so large, .92 and .90, respectively, analyses of variance were computed for both groups using only Days 2 and 3 to insure that performance in Day 1 was not primarily responsible for the significant Days × *S*s interaction obtained with the data for all 3 days. It was found that the Days × *S*s interaction, with Day 1 omitted, was still significant for both groups ($p < .001$) even though the magnitude of the *F* ratio was reduced.

The *S*s' performances on Days 2 and 3 are presented graphically in Figures 2 and 3 for Groups I and II, respectively. Visual inspection of these graphs indicates that the primary reason for the statistically significant differences obtained between Days 2 and 3 is that variability of the C-D pace is so small. Thus with very low variability even small differences between mean C-D pace levels on Days 2 and 3 are found to be statistically significant. For the majority of *S*s in both groups, the C-D pace on Day 2 was essentially the same as Day 3.

These results reveal that even though *S*s' C-D pace on Day 2 may be significantly different statistically from their C-D pace on Day 3 (the significant Days × *S*s interaction), the C-D pace on Day 2 is highly similar to the C-D pace on Day 3 (the significant coefficients of correlation).

The discrepancy between the two *S* groups relative to the correlation of C-D pace between Days 1 and 3, combined with the observation that the overall mean C-D pace for Group I was almost .5 mph lower than Group

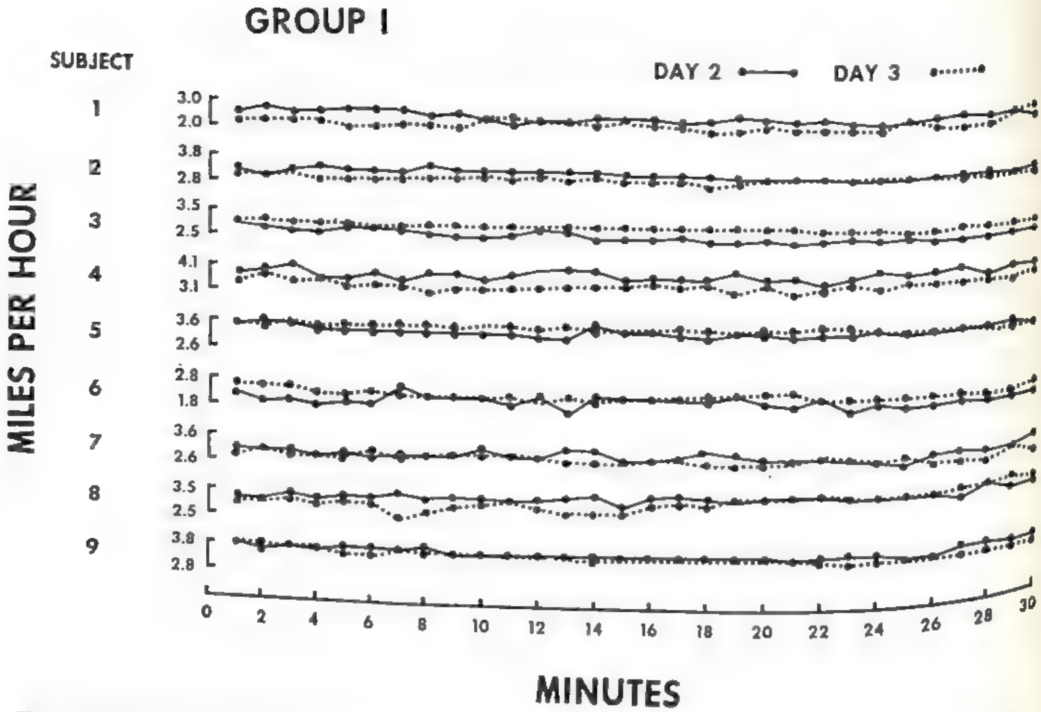


FIG. 2. Mean mph/min for the 30-min. characteristic C-D walk on Days 2 and 3. (Group I)

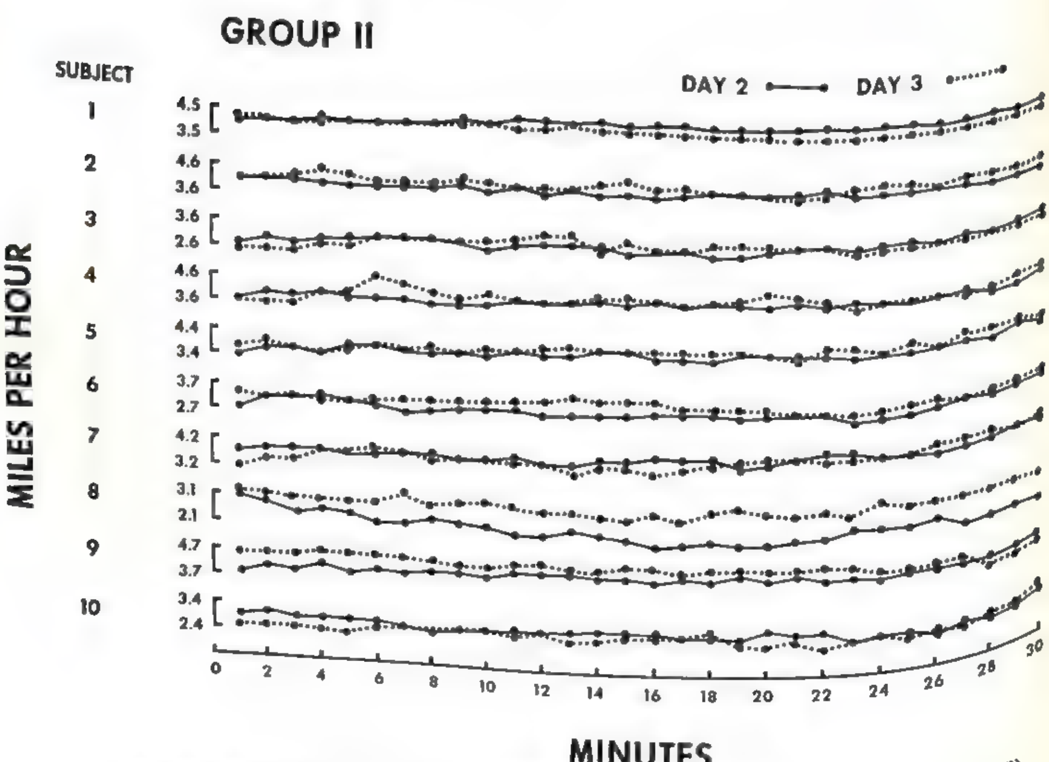


FIG. 3. Mean mph/min for the 30-min. characteristic C-D walk on Days 2 and 3. (Group II)

II, suggests that in measures of this type of performance, large groups of *S*s are to be desired. It is doubtful that the longer pre-C-D pace walks for Group II (8 min. for both the slow to C-D and the fast to C-D walk for Group II as compared to 4 min. for each of the pre-C-D walks for Group I) account for the difference observed. One difference between the two groups of *S*s that should be noted is that most of the *S*s in Group I were next going to paratrooper training whereas none of the *S*s in Group II was scheduled for this training. The extent to which this factor along with other factors (psychological and physiological) were related to differences observed will be the subject of future research.

The next step in this line of research is to determine the effect of initiating a walk with a fast pace (5 mph) versus a slow starting pace (2 mph) in terms of the C-D pace *S*s attain. Based on observations made during the current research, it is hypothesized that only 5–10 min. are required after initiating a walk, either fast or slow, for an *S* to reach his daily characteristic C-D pace.

As the hypothesis that individuals have a characteristic C-D pace that is relatively stable on any given day was supported, it will now be possible to apply the relative loading technique to this dynamic muscle group activity to obtain various measures of endurance. It is hypothesized that of the various possible pace rates or walking measures to be related to various physical, physiological, or psychological measures available, two will be found to be the most meaningful in future research: (*a*) *S*s' variance of the C-D pace on any given day; and (*b*) the variance of *S*s' C-D pace from day to day (Holmgren, 1966).

REFERENCES

BARANY, J. W. Use of a force-platform for the study of hemiplegic gait. Paper read at American Society for Engineering Education, June 1964.

BARANY, J. W., & GREENE, J. H. The force-platform: Instrument for selecting and training employees. *American Journal of Psychology*, 1961, 74, 121–124.

CALDWELL, L. S. Relative muscle loading and endurance. *Journal of Engineering Psychology*, 1963, 2, 155–161.

CALDWELL, L. S. The load-endurance relationship for a static manual response. *Human Factors*, 1964, 6, 71–79.

ELBEL, E. R. Relationship between leg strength, leg endurance and other body measures. *Journal of Applied Physiology*, 1949, 2, 197–207.

EVANS, W. O. The effect of treadmill grade on performance decrement using a titration schedule. USA MRL Rep., 1962, No. 535. (Fort Knox, Kentucky)

EVANS, W. O. A titration schedule on a treadmill. *Journal of the Experimental Analysis of Behavior*, 1963, 6, 219–221.

HOLMGREN, G. L. Characteristic pace—a potential tool for vascular surgeons. USA MRL Rep., 1966, No. 687. (Fort Knox, Kentucky)

ISMAIL, A. H. An operational criterion for physical fitness tests. Unpublished manuscript, Purdue University, 1964.

ROHMERT, W. Ermittlung von Erholungs pau sen für statische Arbeit des Menschen. *Arbeitsphysiologie*, 1960, 18, 123–164.

SMITH, K. U., MCDERMID, C. D., & SHIDEMAN, F. E. Analysis of the temporal components of motion in human gait. *American Journal of Physical Medicine*, 1960, 39, 142–151.

TUTTLE, W. W., JANNEY, L. D., & THOMPSON, C. W. Relation of maximum grip strength to grip strength endurance. *Journal of Applied Physiology*, 1950, 2, 663–670.

(Received June 30, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 284–290

ALIENATION, ENVIRONMENTAL CHARACTERISTICS, AND WORKER RESPONSES

MILTON R. BLOOD AND CHARLES L. HULIN
University of Illinois, Urbana

Data gathered from 1,900 male workers located in 21 plants in the eastern United States are presented. These data are analyzed to determine the influence of environmental characteristics presumed to index feelings of alienation from middle-class norms. Predictions were made that workers in communities which should foster integration with middle-class norms would structure their jobs differently and would respond differently than alienated workers. Workers in communities fostering integration with middle-class norms should report higher satisfaction on highly skilled jobs. They should value retirement and should plan for it while working. Alienated workers should report lower satisfaction on highly skilled jobs. Pay should have a stronger effect on the satisfaction of alienated workers, and these workers would be more likely to look for other work after retirement. The predictions were regarded as confirmed for blue-collar workers. The implications of these findings for striving need-theoretic models of human motivation are discussed.

Some recent industrial field studies have pointed up the importance of community variables as determiners of workers' responses. Katzell, Barrett, and Parker (1961) and Cureton and Katzell (1962) found job satisfaction and performance inversely related to the degree of urbanization of the community. They attribute this relationship to differences in needs and expectancies of the workers in the various environments. With increased urbanization, needs and expectancies rise and there is less satisfaction from any specific return. Hulin (1966), using the worker's frame of reference as an intervening variable, predicted and empirically verified that job satisfaction is higher in communities with substantial slum areas. The assumption of his discussion is that the worker assesses his present status by referring to the alternative positions which are available to him. Since attractive alternatives are not readily apparent in slum conditions, the worker's present job will be seen as relatively more satisfactory. Turner and Lawrence (1965), in a study of workers' responses to the technological aspects of the work situation, found that rural and small town workers were more satisfied when their jobs were more autonomous, required more skill, were more varied, and contained more social interaction and responsibility. In essence, the most satisfying jobs demanded greater personal involvement. This was, of course, the "expected" result. In contrast to this, city workers were more likely to be satisfied when their jobs were less personally involving. One of the possible explanations offered by Turner and Lawrence for this unexpected response from city workers is the notion of anomie, a state of societal normlessness brought about by industrialization, which has been frequently investigated in recent sociological researches.

In addition to these studies there are a growing number of studies of "job enlargement" which present conflicting results. In general, results in line with predictions are obtained if the studies are done on workers with rural backgrounds, but contrary results are obtained if the workers are from urban backgrounds (see Friedlander, 1965; Kennedy & O'Neill, 1958; Kilbridge, 1960).

From these studies and from sociological studies of anomie and alienation a construct can be formulated which can be used in structuring and predicting workers' responses. This construct might most efficiently be conceived as a continuum running from "integration with middle-class norms" to "alienation from middle-class norms." However, it should be pointed out that this is a complex phenomenon described as unidimensional only for ease of conceptualization. At the integrated end of the construct are found workers who have personal involvement with their jobs and

aspirations within their occupations. Their goals are the type of upward mobility, social climbing goals generally associated with the American middle class. At the opposite pole of the construct, workers can be described as involved in their jobs only instrumentally; that is, the job is only a provider of means for pursuing extraoccupational goals. The concern of these workers is not for increased responsibility, higher status, or more autonomy. They want money, and they want it in return for a minimal amount of personal involvement. This difference between integrated and alienated workers is similar to Dalton's (1947) discrimination between overproducers who are likely to hold middle-class aspirations and underproducers who do not identify with middle-class ideals. The construct of alienation being proposed in this study stands in obvious relation to the Protestant ethic proposed and discussed by Weber (1958). It is likely that conditions fostering integration with middle-class norms will also foster adherence to the Protestant ethic since the latter is an aspect of the former.

What environmental conditions should lead to alienation from middle-class norms? Data have suggested that anomie is associated with lower-class highly industrialized situations (Bell, 1957; Dean, 1961; Killian & Grigg, 1962; Mayo, 1933; McClosky & Schaar, 1965; Ruitenbeck, 1964; Simpson & Miller, 1963; Turner & Lawrence, 1965). Though anomie is different conceptually from the construct being defined here, the measurement of anomie has been such that it might just as easily be interpreted as alienation from the middle class. It is postulated that "alienation from middle-class norms" results from lack of socialization to middle-class norms. That is, where a segment of society exists which holds non-middle-class norms and which is large enough to sustain its own norms, the members of that subculture will become socialized to the norms of that subculture. A handful of industrial workers in a small community could not be expected to sustain a separate set of norms, but persons separated from middle-class identification by low educational attainment or low occupational status and living in ghettos, slums, and highly industrialized communities could develop and sustain a distinct norm system. Alienation from middle-class norms, then, is fostered by industrialized, socially heterogeneous, metropolitan conditions. This conceptualization of the background of the alienated worker coincides with the characterizations by Whyte (1955) of restrictors and rate busters. The social and family background of restrictors was urban, working class, whereas rate busters were from farms or lower-middle-class families. Worthy (1950) has also pointed out the possible effects of living in urban areas on the motivation and especially the morale of industrial workers.

Method

Before describing the methods of this investigation, two cautions must be invoked. First, though the data to be presented are not such that they could support the contention of causality, a causal discussion is used here. A causal theory in this sense serves only as a working hypothesis to be altered as the data demand. This does not detract, however, from the usefulness of the construct as a guide to research and application. Second, no value judgment is intended by the use of the word "alienated." It would be as logical to consider such workers integrated to their norms from which middle-class persons are alienated (and whether or not it would be better for the workers or for society for them to be integrated to middle-class norms is a social question beyond the proper boundaries of this paper). Orientation to middle-class norms has been adopted as the point of reference here because middle-class norms predominate in our culture and among social scientists.

Data for this study were provided from a study carried out at Cornell University (Smith, Kendall, & Hulin, in press). Subjects (*Ss*) were 1,390 male blue-collar workers and 511 male white-collar workers representing 21 different plants located throughout the eastern half of the United States. These data were gathered in 1961 and 1962 as part of a large-scale study of retirement satisfaction sponsored by the Ford Foundation.[1]

Stimulus Variables

To predict from the alienation construct it was necessary to have stimulus variables which indexed the environmental conditions which foster alienation and another set of variables which indicated the responses of workers to their jobs. The variates used to index community characteristics which should foster alienation among the workers were chosen from a principal component analysis of an intercorrelation matrix of per capita census variables originally provided by Kendall (1963). While the origi-

[1] The authors would like to thank Patricia Cain Smith of Bowling Green University who very generously made these data available for this analysis.

TABLE 1

WORKER RESPONSE VARIABLES AND THE CONDITION UNDER WHICH THE VARIABLE IS PREDICTED TO BE HIGHER

Variable	Alienated	Integrated
1. Importance of planning for retirement		x
2. Made plans yet for retirement		x
3. Preparation for retirement index		x
4. Look for other work after retirement	x	
5. Personal satisfaction from the job		x
6. JDI work satisfaction-Skill correlation		x
7. JDI work satisfaction-Job-level correlation		x
8. JDI work satisfaction-JIG correlation		x
9. JDI pay satisfaction-JIG correlation		x
10. JDI promotion satisfaction-JIG correlation	x	
11. JDI work satisfaction-LIG correlation		x
12. JDI pay satisfaction-LIG correlation		x
13. JDI promotion satisfaction-LIG correlation	x	
14. JIG-LIG correlation		x

nal solution was simply a set of statistical variates which could be used to describe communities, several of these dimensions appear to be useful as indexes of alienation. These variates consisted of the weighted sum of variables where the weights were proportional to the loadings of the variables on the component. The complex variates were used since it was felt that the variates would provide more reliable and meaningful indexes than would individual variables. See Kendall (1963) for a complete description of this solution.

The variates used in this research were descriptively named *Slum Conditions* indexed mainly by the weighted sum of the standard scores of percentage of native white (reversed scoring), percentage of nonwhite, and percentage of owner-occupied housing (reversed scoring); *Urbanization* indexed mainly by the sum of percentage of rural nonfarm (reversed scoring), percentage of urban population, total population, and per capita motor vehicle deaths (reversed scoring); *Urban Growth* indexed mainly by percentage of immigration, percentage of dwellings vacant (reversed scoring), and percentage of new homes; *Prosperity and Cost of Living* indexed by the sum of percentage of sound housing, medium income, percentage of workers in wholesale, per capita retail sales, and percentage with income over $10,000; and *Productive Farming* indexed by the sum of percentage of workers in wholesale (reversed scoring), average farm income, percentage of workers in agriculture, and percentage of change in farm level of living. A sixth variable *Population Density* (population per square mile) was chosen as a final index of alienating conditions.

There were no data indicating that these variates behave in the manner predicted. Thus, confirmation of the predictions of this report would serve also to enlarge existing knowledge of the community conditions fostering alienation. It was possible that one or more of the variates would prove to be a poor index for the present purpose, and it was almost certain that these six indexes would differ in the strength with which they gauge the postulated conditions. For this reason, all of them were used in the position of the independent variable. Even if some of them did not show the anticipated relationship or showed it only weakly, the consideration of all six of them together provided more understanding of the construct (cf. Webb, Campbell, Schwartz, & Sechrest, 1966).

Response Variables

Fourteen variables were chosen which were expected to show differences between integrated and alienated workers. Four of them concerned retirement. Workers were asked to rate the importance of planning for retirement, whether or not they had made plans yet for retirement, and whether they would look for other work after they retired. Also a Preparation for Retirement Index was established for each worker indicating the extent of preparation the worker had made for his retirement years. On the three planning-and-preparation-for-retirement variables it was predicted that the integrated workers would score higher since it was felt that the desire for a leisurely retirement has become imbued with a great deal of prestige or status significance in the middle class. For the same reason, integrated workers were expected to be less likely to say that they would look for other work after retirement.

Workers were asked to rank their job as well as their family life, hobbies, etc., as a provider of personal satisfaction. Integrated workers were expected to be more likely to rank their job first or second as a provider of personal satisfaction.

It was predicted that the correlation of the work satisfaction scale of the Job Description Index (JDI) with a rating of job level would be lower for alienated workers than for integrated workers since alienated workers are more satisfied when their jobs are *less demanding*. The same prediction was made for the correlation of the JDI work satisfaction scale and ratings of job-skill requirements.

As the job was expected to play a more central role in the lives of the integrated workers, the correlation between satisfaction with the job in general (JIG) and satisfaction with life in general (LIG) was predicted to be higher for them. JIG and LIG Scales were measured by General Motors Faces Scales (Kunin, 1955). Using JDI scales, satisfaction with

TABLE 2

PREDICTION TESTS USING ALL WORKERS: CORRELATIONS (PEARSON r'S) BETWEEN STIMULUS AND RESPONSE VARIABLES

Response variables	Predicted direction	Stimulus variables: Slum Conditions	Urbanization	Urban Growth	Cost of Living	Productive Farming (reversed)	Population Density
Importance of planning for retirement	−	−.05	−.46	+.31	−.42	−.16	−.51
Made plans yet for retirement	−	−.18	−.60	+.24	−.57	−.20	−.61
Preparation for retirement index	−	−.37	−.33	+.15	−.28	−.01	−.41
Look for other work after retirement	+	+.35	+.11	+.21	+.18	−.36	+.25
Personal satisfaction from the job	−	+.50	+.16	+.43	+.19	−.15	00
Correlations between:							
Work satisfaction-Skill level	−	00	−.03	−.07	−.05	+.07	+.06
Work satisfaction-Job level	−	−.25	−.23	−.14	−.19	−.24	−.13
Work satisfaction-Job-in-general satisfaction	−	+.26	+.31	−.04	+.33	+.32	+.38
Pay satisfaction-Job-in-general satisfaction	+	−.19	−.02	−.06	+.04	+.19	+.09
Promotion satisfaction-Job-in-general satisfaction	−	−.31	−.15	−.38	−.19	+.18	−.04
Work satisfaction-Life-in-general satisfaction	−	+.39	+.32	+.08	+.30	+.27	+.34
Pay satisfaction-Life-in-general satisfaction	+	−.28	00	−.21	−.01	+.28	+.28
Promotion satisfaction-Life-in-general satisfaction	−	00	−.05	−.20	−.13	+.28	+.04
Job-in-general satisfaction-Life-in-general satisfaction	−	+.16	+.14	−.11	+.07	+.19	+.02

Note.—N = 21 companies. Bold-faced numbers are in the direction of the prediction.

work and satisfaction with promotional opportunities should be more highly correlated with JIG and LIG in the integrated sample. On the other hand, pay satisfaction should be more highly correlated with JIG and LIG in the alienated sample because these are the persons who view their job as primarily an activity which is instrumental to the achievement of other goals. A summary of these predictions can be seen in Table 1 where the response variables are listed and an "x" is placed under the heading of the situation in which the level of the variable was predicted to be higher.

It was possible to assign each of the 21 plant locations a score on each of the indexes of alienating conditions and on each of the indexes of workers' responses. The analysis was carried out by correlating these two sets of variables—the alienation indexes and the response variables. It should be noted that the n for each of these correlations was 21 and not 1,900. This of course means that this study, in spite of the large number of subjects, was a very small study considering the type of analysis which was being carried out. Further, 54 of the predictions are predictions about the relative size of a *correlation*. Differences were being predicted in the relationship between two response variables as determined by differences in the stimulus condition. For these two reasons a prediction will be simply regarded as being supported if the finding was in the expected direction and the results will be discussed as a whole and not as specific findings.

RESULTS

Using all of the workers of the sample, the correlations were computed and the results can be seen in Table 2. The first column gives the predicted direction of the correlations for each row. Of the 84 predictions represented in this table, 45 are in the predicted direction. This, of course, does not support the construct.

It scarcely seemed appropriate to abandon the construct after such limited analysis, and a further step was proposed. It was assumed that middle-class ideals were much more likely to be found among white-collar workers than among blue-collar workers, regardless of community environment. If that assumption is correct, the results might have been masked

TABLE 3

PREDICTION TESTS USING BLUE-COLLAR WORKERS: CORRELATIONS (PEARSON r'S) BETWEEN STIMULUS AND RESPONSE VARIABLES

Response variables	Predicted direction	Stimulus variables					
		Slum Conditions	Urbanization	Urban Growth	Cost of Living	Productive Farming (reversed)	Population Density
Importance of planning for retirement	−	+.23	**−.22**	+.42	**−.18**	**−.09**	**−.35**
Made plans yet for retirement	−	**−.25**	**−.55**	+.09	**−.57**	**−.05**	**−.65**
Preparation for retirement index	−	**−.46**	**−.19**	**−.11**	**−.20**	+.20	**−.28**
Look for other work after retirement	+	**+.32**	**+.11**	**+.02**	**+.16**	−.22	**+.23**
Personal satisfaction from the job	−	+.52	+.16	+.46	+.20	**−.21**	+.01
Correlations between:							
Work satisfaction–Skill level	−	**−.03**	**−.07**	**−.10**	**−.09**	**−.02**	+.07
Work satisfaction–Job level	−	**−.65**	**−.18**	**−.47**	**−.20**	+.17	**−.14**
Work satisfaction–Job-in-general satisfaction	−	**−.12**	+.20	**−.34**	+.17	+.37	+.25
Pay satisfaction–Job-in-general satisfaction	+	**+.13**	**+.02**	**+.18**	**+.11**	**+.09**	**+.12**
Promotion satisfaction–Job-in-general satisfaction	−	**−.31**	**−.23**	**−.30**	**−.23**	**−.02**	**−.17**
Work satisfaction–Life-in-general satisfaction	−	**−.31**	+.15	**−.48**	+.05	+.47	+.13
Pay satisfaction–Life-in-general satisfaction	+	**+.05**	**+.13**	**+.09**	**+.10**	**+.37**	**+.28**
Promotion satisfaction–Life-in-general satisfaction	−	**−.48**	**−.10**	**−.47**	**−.18**	+.26	**−.06**
Job-in-general satisfaction–Life-in-general satisfaction	−	**−.30**	+.03	**−.22**	**−.01**	.00	**−.21**

Note.—N = 21 companies. Bold-faced numbers are in the direction of the prediction.

by the inclusion of white-collar workers in the sample. Accordingly, the 511 white-collar workers were dropped, and the analyses were redone. Again showing the direction of predictions in the first column, the results using only blue-collar workers are shown in Table 3. Here 61 of 84 predictions are in the proper direction. Using the normal approximation to the binomial distribution, the probability of finding this many confirmations out of 84 predictions by chance is less than .001. Though extreme caution must be used in attempting to interpret the results of any specific variable, three of the response variables warrant mention. Those are the three which did *not* confirm the construct. JDI work–JIG correlation, and JDI work–LIG correlation do not present any clear pattern. The equivocal results with these two correlations are not easily explained. Perhaps the relationship between work satisfaction and global satisfaction is too strong to allow modification by the social environmental differences which are being measured here. The other nonconfirming variable, "Personal satisfaction from the job," is in the direction opposite the prediction in five of the six opportunities. The present analysis is not so powerful that this response can be concluded to act in this direction in all cases. This particular response and related areas should certainly enjoy increased interest in any continuation of this line of research.

The Slum Conditions, Cost of Living, and Urban Growth variates work best as predictors of alienated responses having 12, 11, and 11 correct predictions, respectively. Urbanization and Population Density each had 10 correct predictions, while Productive Farming had only 7. Thus, empirically as well as logically, the Slum Conditions variate is the one most closely aligned with the alienation construct.

The Productive Farming variate does not seem useful.

DISCUSSION

In general it appears that the construct of alienation has been demonstrated to be useful for structuring workers' responses. It has been shown that workers living in communities which should foster alienation from middle-class norms structure their jobs and their lives predictably differently from workers in communities where adherence to middle-class norms would be expected. Further, these differences were predicted from a theoretical framework consistent with the construct of alienation. Since this study should be regarded as a pilot study, this discussion will emphasize the problems and future research direction rather than the implications of the findings.

It is true that in this study as in the previous research, "alienation from middle-class norms" is not the only tenable hypothesis. Katzell et al. (1961) invoked an explanation based on differing "needs" being generated by urban and rural environments. Hulin (1966) and Kendall (1963) used the concepts of "frame of reference" and the "alternatives available to the workers" in economically depressed and slum-ridden communities. Turner and Lawrence (1965) used the concept of "anomie" to explain their results. Even in the present study it is possible that an economic-frame-of-reference hypothesis would be useful in understanding the results. The cost of living variate made 11 out of 14 correct predictions suggesting that frames of reference of the workers in the community play a significant role in structuring workers' jobs. On the other hand, communities with a high cost of living would be expected to be less than optimal as places for workers to live. Alienation from the norms of the dominant social group might be a reasonable response of the economically disadvantaged workers who might well see the "middle class" as the cause of their unenviable position. Thus, both of these constructs may be similar. Also, since "productive farming" proved useful in predicting level of pay satisfaction consistent with a frame-of-reference explanation (Hulin, 1966; Kendall, 1963) but not alienated responses regarding job structure, the economic-frame-of-reference explanation in the present study is a bit tenuous. It also could be argued that anomie remains a tenable explanation for these findings and all that has been found is that "integrated" workers are more predictable than "alienated" workers who are truly normless (anomic). Table 1 indicates that in the majority of the cases prediction has been made for stronger relationships between response variables for the integrated sample. However, of the 18 cases where higher relationships were predicted for the alienated workers, 17 were in the expected direction. Thus, it does not appear that anomie is a tenable explanation. The alienated workers are not normless. They have norms but they are different from those of the middle classes. It would appear that "alienation from middle-class norms" is the only construct so far invoked which is capable of structuring all of these diverse findings.

The construct was not confirmed above the level of skilled blue-collar workers. The fact that it was not found in a sample of white-collar workers is in line with the definition of the construct. The possibility that different occupations may generate different susceptibility to alienation is similar to Blauner's (1964) reasoning and certainly has not been ruled out. While the upper boundary of blue-collar workers was arbitrarily set at the level of skilled workers in this study, future research may show that some other level will be optimal in applications of the construct.

The present data delineate attitude changes which accompany alienation. An area of at least equal and probably greater importance will be to see if there are related behavioral changes. Related to job-performance differences is the question of what job-design criteria will be most effective in maximizing satisfaction and performance among alienated workers. As the Turner and Lawrence (1965) findings suggest, the best job design in alienating conditions may be contrary to the models usually proposed by human-relations-oriented investigators. Although integrated workers desire greater responsibility and autonomy, alienated workers may be happiest when given a job which demands little personal involvement either in terms of task skills or identification with the goals of management.

It should be reiterated that this represents a pilot study and although predictions regarding all of the relationships were made a priori there is still only a rudimentary knowledge of the effects of the independent variable. Eventually, with increasing knowledge and sophistication one may know enough about alienation to attempt direct psychological assessment rather than indexing it by means of environmental variables. To the deplorers of "actuarial research" (such as has been conducted here) this would represent a significant step forward. However, if this direct assessment is carried out, researchers will be back in the domain of verbal response–verbal response correlations with all of the attendant problems of response sets, halo, acquiescence, and unreliability. Even with the "slippage" that occurs between an index and a construct, it seems preferable to use a stimulus-response paradigm as was employed in this study.

This study does seem to have strong implications for theorists who would like to talk about "basic human needs." Both the currently popular need hierarchy approach and the human relations theorists talk about needs for self-actualization, needs for autonomy, needs for a demanding job, etc., which are supposedly basic needs of all people. These data indicate that at the very least these systems need to be revised to take cultural differences into account.

REFERENCES

BELL, W. Anomie, social isolation, and the class structure. *Sociometry*, 1957, **20**, 105–116.

BLAUNER, R. *Alienation and freedom: The factory worker and his industry*. Chicago: University of Chicago Press, 1964.

CURETON, E. E., & KATZELL, R. A. A further analysis of the relations among job performance and situational variables. *Journal of Applied Psychology*, 1962, **46**, 230.

DALTON, M. Worker response and social background. *Journal of Political Economy*, 1947, **55**, 323–332.

DEAN, D. G. Alienation: Its meaning and measurement. *American Sociological Review*, 1961, **26**, 753–758.

FRIEDLANDER, F. Comparative work value systems. *Personnel Psychology*, 1965, **18**, 1–20.

HULIN, C. L. Effects of community characteristics on measures of job satisfaction. *Journal of Applied Psychology*, 1966, **50**, 185–192.

KATZELL, R. A., BARRETT, R. S., & PARKER, T. C. Job satisfaction, job performance, and situational characteristics. *Journal of Applied Psychology*, 1961, **45**, 65–72.

KENDALL, L. M. Canonical analysis of job satisfaction and behavioral, personal background, and situational data. Unpublished doctoral dissertation, Cornell University, 1963.

KENNEDY, J. E., & O'NEILL, H. E. Job content and workers' opinions. *Journal of Applied Psychology*, 1958, **42**, 372–375.

KILBRIDGE, M. D. Do workers prefer larger jobs? *Personnel*, 1960, **37**(5), 45–48.

KILLIAN, L. M., & GRIGG, C. M. Urbanism, race, and anomia. *American Journal of Sociology*, 1962, **67**, 661–665.

KUNIN, T. The construction of a new type of attitude measure. *Personnel Psychology*, 1955, **8**, 65–77.

MAYO, E. *The human problems of an industrial civilization*. New York: Macmillan, 1933.

MCCLOSKY, H., & SCHAAR, J. H. Psychological dimensions of anomy. *American Sociological Review*, 1965, **30**, 14–40.

RUITENBECK, H. M. *The individual and the crowd*. New York: Nelson, 1964.

SIMPSON, R. L., & MILLER, H. M. Social status and anomia. *Social Problems*, 1963, **10**, 256–264.

SMITH, P. C., KENDALL, L. M., & HULIN, C. L. *Measurement of satisfaction in work and retirement*. New York: Rand McNally, in press.

TURNER, A. N., & LAWRENCE, P. R. *Industrial jobs and the worker: An investigation of response to task attributes*. Boston: Harvard University Press, 1965.

WEBB, B. J., CAMPBELL, D. T., SCHWARTZ, R. D., & SECHREST, L. *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally, 1966.

WEBER, M. *The protestant ethic and the spirit of capitalism*. (Tr. by Talcott Parsons) New York: Scribner, 1958.

WHYTE, W. F. *Money and motivation: An analysis of incentives*. New York: Harper, 1955.

WORTHY, J. C. Organizational structure and employee morale. *American Sociological Review*, 1950, **15**, 169–179.

(Received July 5, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 291–294

DETERMINANTS OF WORK ATTITUDES AMONG NEGROES[1]

ROBERT BLOOM AND JOHN R. BARRY[2]

University of Florida

The motivator-hygiene theory of work attitudes assumes that 2 independent sets of variables determine job satisfaction and job dissatisfaction. The generality of this theory was tested with responses from 85 Negro blue-collar workers to a 40-item questionnaire. These data were factor analyzed and compared with comparable data from 117 white blue-collar workers. The 2 sets of data appeared to differ although some similarities were found. It appeared that hygiene factors were more important to the Negroes than to the whites. The 2-factor theory may be too simple to encompass the concepts of job satisfaction and dissatisfaction, at least in the blue-collar Negro. It may be that the 2-factor theory is less useful when one considers low-status work.

Most of the research on the two-factor motivator-hygiene theory (Herzberg, Mausner, & Snyderman, 1959) has involved white-collar workers. Friedlander (1963, 1964, 1965), testing engineering supervisors, salaried employees, scientists, and several other occupational groups with a variety of questionnaires and tests, concluded that, although motivation and hygiene factors are important in job satisfaction and dissatisfaction, they are not independent as would have been predicted from Herzberg's theory. Ewen (1964), discussing the motivator-hygiene theory, states:

> A more extensive research design is necessary in order to adequately test the Herzberg theory. For the present, however, it must be concluded that the nature of satisfiers and dissatisfiers (if such variables do in fact exist) is as yet far from clear, and may be different in different jobs [p. 163].

Wernimont (1966), studying accountants and engineers (as did Herzberg et al.), elicited responses to both forced-choice and free-choice items to describe past satisfying and dissatisfying job situations. He concluded that both motivator and hygiene factors contributed to satisfaction and dissatisfaction. Malinovsky and Barry (1965) designed a 40-item Work Attitude Survey (WAS) to measure work attitudes of blue-collar workers. The WAS consisted of 20 motivator and 20 hygiene items which were rated on a Likert-type 5-point scale. There were 4 items representing each one of the 10 work-attitude variables taken from the motivator-hygiene theory of job satisfaction. The sample consisted of 117 white blue-collar workers employed in the grounds crew (primarily maintenance men and watchmen) in a large Southern state university.

Malinovsky and Barry obtained 12 factors which are listed in Table 1. For the most part, these factors were composed of either motivator or hygiene items. A second-order factor analysis yielded two general factors, which roughly corresponded to the motivator and the hygiene sets of variables. However, the authors noted that these two sets of work-attitude variables were *not* as completely independent of each other in this study as might be predicted from Herzberg's theory.

The present study was designed to determine whether work attitudes of Negro blue-collar workers could be explained by Herzberg's two-factor theory. Also, a comparison with the Malinovsky-Barry data from white blue-collar workers was planned.

METHOD

The WAS form which had been designed to study the work attitudes of white blue-collar workers by Malinovsky and Barry (1965) was used in this present study. The WAS was distributed to 180 Negro, male, blue-collar workers in June 1964, at the same time that it was distributed to the white

[1] This research was supported in part by a research grant, RD-1127, from the Vocational Rehabilitation Administration, Department of Health, Education and Welfare, Washington, D. C. The authors also wish to thank M. R. Malinovsky and G. H. Dunteman for help in data collection and analysis, respectively.

[2] From the Regional Rehabilitation Research Institute, College of Health Related Professions.

TABLE 1

LISTING OF FACTORS AND COMMON ITEMS IN THE 2 STUDIES

Bloom-Barry study		Malinovsky-Barry study		
Factor name	A[a]	Factor name	A[a]	B[b]
I. Work Situation	5	IX. Work Frustrations	5	3
II. External Guidance	7	VII. Physical Work Environment	9	5
III. External Incentive	5	VII. Physical Work Environment	9	3
IV. Salary	3	I. Salary	4	2
V. Advancement	3	IV. Advancement	3	2
VI. External Encouragement	5	VIII. Unrecognized Work Efforts	3	2
VII. Work Satisfactions	5	II. Technical Supervision	2	—[c]
VIII. Supervisory Relations	3	III. Interpersonal Relations	2	—[c]
IX. No Interpretation	2	V. Individual Accomplishments	2	—[c]
X. No Interpretation	3	VI. Work Role Dissatisfactions	2	—[c]
XI. No Interpretation	2	X. No Interpretation	2	—[c]
XII. Peer Relations	3	XI. No Interpretation	2	—[c]
XIII. Work Frustration	3	XII. No Interpretation	3	—[c]

[a] A = Total number of items loading over .40.
[b] B = Number of common items in same factor in the 2 studies.
[c] Factors not considered comparable.

blue-collar workers in the Malinovsky-Barry study. Eighty-five questionnaires (47% as compared to 43% from the white workers group) were completed and returned anonymously by the Negroes. These 85 Negro workers comprised the sample in the present study. All 180 workers were employed by the Plants and Grounds Department in a large Southern state university. Of the 85 returnees, approximately 50% were employed as janitors and the rest in various other unskilled and semiskilled jobs. In the white sample, less than 5% were employed as janitors and the majority of the rest in maintenance positions at all skill levels.

The age of the Negro workers ranged from 18 to 63 yr. with a mean age of 43 yr. The mean age in the white sample was 41 yr. with a range of 21–68 yr. The white group's yearly base salary ranged from $2,700 to $5,900 with a mean yearly salary of about $3,700. The Negroes yearly average was about $3,000 with a range from less than $2,700 to $5,000. The length of employment at the university for the Negro group ranged from 1 mo. to 19 yr. with a mean length of employment of 6 yr. The white workers' length of employment ranged from 1 mo. to 27 yr. with a mean length of employment of 5.5 yr.

Pearson product-moment correlation coefficients were computed among the 40 motivator-hygiene items and the resultant 40×40 intercorrelation matrix was factor analyzed by the principal factor method and machine rotated by the Varimax routine. Unities were used in the principal diagonals. Only those factors with eigenvalues of 1 or above were rotated.

RESULTS

The 13 factors which were extracted from the rotated factor matrix are listed in Table 1. All items with loadings of .40 or above were considered as contributing to that factor. Of the 13 factors extracted, 8 factors contained both motivator and hygiene items while 2 (Factors II and IX) contained only hygiene items, and 3 (Factors X, XI, and XIII) contained only motivator items. In general, the hygiene items had higher loadings and determined more of the variance in the factor matrix than the motivator items.

Factor I—The Work Situation—appeared to reflect specific feelings toward many aspects of the work context. This factor accounted for approximately 10% of the common variance. Subsequent factors, II, III . . . XII, accounted for decreasing amounts of the variance, Factor XIII accounting for only about 5% of the variance. Factor II—External Guidance—was composed of items suggesting that the worker felt that his job environment was controlled by supervisors and other company decisions. Factor III—External Incentive—reflected the workers' concern with the job content, rather than the interpersonal element. Some of the items contributing to this factor suggested that the type of equipment used by the worker was important to him as an indication of status.

Factor IV was determined by Salary items and Factor V was concerned with worker Advancement. Factor VI—External Encouragement—reflected the importance of encouragement and recognition by supervisors. Factor VII—Work Satisfactions—reflected worker pride and satisfaction in the intrinsic aspects of the work itself. Factor VIII—Supervisory Relations—reflected the importance of personal relations with those in a higher job status. Factor IX, X, and XI appeared to be uninterpretable. Factor XII—Peer Relations—reflected the importance of peer relations in the work context. Factor XIII—Work Frustration—suggested the need of the worker to feel responsible and independent.

A comparison of the factors and the number of items common to each factor in this study and that of Malinovsky and Barry

(1965) is shown in Table 1. Substantial differences as well as some similarities are evident in the two studies. The two sets of data are comparable in that Factors I, II, III, IV, V, and VI in the present study are roughly similar to Factors IX, VII, I, IV, and VIII, respectively, in the Malinovsky and Barry study. On the other hand, the fact that most of the factors in the present study contained both motivator and hygiene items suggests that the Herzberg two-factor theory is not as well substantiated as in the Malinovsky-Barry study. Furthermore, the present findings show a far more complicated picture of the work motivations of the Negro than would have been predicted from that theory.

DISCUSSION

The implications of this study regarding the attitudes of the Negro blue-collar worker are somewhat attenuated in that only 47% of the WASs were returned, and a few of those were not completely filled out. A question also might be raised as to whether the Negroes perceived the items on the WAS in the same way as the whites did. Probably they did not, since the item overlap in the factors from the studies was relatively small (see Table 1). From this disparity in perception, and from the content of the items in the factors, it can be assumed that, if in fact, motivator and hygiene variables do exist in the Negro frame of reference, hygiene variables are more important to the Negro than are motivator variables.

Friedlander (1966), testing both white- and blue-collar workers, found that for the blue-collar worker there were no significant relationships between any of the motivational measures used and job performance. Friedlander further described a hierarchy of factors important to the blue-collar worker with his factor, Social Environment of Work, heading the list. For the blue-collar group "the pre-work culture evidently forms a motivational system which remains intact throughout, and is probably reinforced by, the work environment [Friedlander, 1966, p. 151]."

Using Hatt and North's (1962) lists of Prestige Ratings of Occupations, the occupation of janitor was found to be rated lower than that of carpenter, machine operator, plumber, etc. The lower job prestige and the lower mean salary of the Negro workers, when considered in relation to other cultural differences between the white and Negro groups, are consistent with the suggestion from this study that hygiene factors are more important to the Negro. For as Herzberg et al. (1959) pointed out, hygiene needs must be met before motivator needs become operative. Perhaps these data represent a stage in the maturation process of a working subclass.

Parker and Kleiner (1966) have described and documented the failure of Negroes to share in the American Dream. While exposed to many of the same stimuli as whites, Negroes often find the pathways to achievement blocked. Data in the present study suggest a difference in the perceptions and goal structures of white and Negro workers. The present study does not imply cause nor effect. The attitudinal differences found in this study may both cause and result from cultural factors, and certainly will limit how these men and their families perceive their futures.

Friedlander (1966) generalized that the cultural disparity between white- and blue-collar workers was sufficiently great that "generalization concerning the motivation-performance relationship cannot be made from one cultural group to the other [p. 151]." Thus the farther one departs from the higher status occupations, the less adequately Herzberg's theory appears to represent the true attitudinal mechanisms.

Certainly a great deal of further research is needed to understand the motivations of the blue-collar worker in general and the Negro blue-collar worker in particular. To date, the diversity of research methods, techniques, and samples probably has been great enough to account for the diverse findings. Whereas subcultural differences undoubtedly affect perception and attitudes, the definitive study relating these differences to work attitudes has yet to be made.

REFERENCES

EWEN, R. B. Some determinants of job satisfaction: A study of the generality of Herzberg's theory. *Journal of Applied Psychology*, 1964, 48, 161–163.

FRIEDLANDER, F. Underlying sources of job satisfaction. *Journal of Applied Psychology,* 1963, **47**, 246–250.

FRIEDLANDER, F. Job characteristics as satisfiers and dissatisfiers. *Journal of Applied Psychology,* 1964, **48**, 388–392.

FRIEDLANDER, F. Relationships between the importance and the satisfaction of various environmental factors. *Journal of Applied Psychology,* 1965, **49**, 160–164.

FRIEDLANDER, F. Motivations to work and organizational performance. *Journal of Applied Psychology,* 1966, **50**, 143–152.

HATT, P. K., & NORTH, C. C. Prestige ratings of occupations. In S. Nosow & W. H. Form (Eds.), *Man, work, and society.* New York: Basic Books, 1962. Pp. 277–283.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. *The motivation to work.* New York: Wiley, 1959.

MALINOVSKY, M. R., & BARRY, J. R. Determinants of work attitudes. *Journal of Applied Psychology,* 1965, **49**, 446–451.

PARKER, S., & KLEINER, R. J. *Mental illness in the urban Negro community.* New York: Free Press of Glencoe, 1966.

WERNIMONT, P. F. Intrinsic and extrinsic factors in job satisfaction. *Journal of Applied Psychology,* 1966, **50**, 41–50.

(Received July 5, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 295-299

THE GRAPHOANALYTIC APPROACH TO SELECTING LIFE INSURANCE SALESMEN

S. M. ZDEP AND H. B. WEAVER[1]

University of Hawaii

Using handwriting samples from 63 life insurance salesmen, an attempt was made to select both successful and unsuccessful salesmen through use of Graphoanalysis. The criterion was based upon 1st-policy-year commissions adjusted for years of experience for each salesman. It was not possible to select successful salesmen by analyzing the writing for 13 personality traits selected intuitively by the analysts, nor was it possible to do this using dominant traits "discovered" in the handwriting of successful salesmen. However, it appears that it may be possible to select failures on the basis of "lacking" certain traits as determined by their handwriting. It is recommended that this aspect be investigated in future validity studies.

Within the past several years there has been increasing interest in the use of handwriting analysis both as a diagnostic and as a selection technique. A summary of research in this area is provided by Fluckiger, Tripp, and Weinberg (1961).

Sonnemann and Kernan (1962) reported successful results using handwriting analysis to select supervisors in an industrial situation. They also pointed out the extensive use of handwriting analysis in European selection of employees, especially in West Germany. Lewinson (1961) stated that the European acceptance of handwriting analysis is greater than in the United States and that it is taught in many universities. Moreover, professional handwriting-analysis organizations exist within the framework of national psychological associations in Holland, France, Switzerland, Germany, and Austria.

In the United States, the prevailing attitude toward the use of handwriting analysis in industry was expressed by Bellows (1961) when he stated that a final decision on its applicability was not yet possible due to the incompleteness of available evidence.

The present investigation was undertaken with the hope of providing some of this evidence. Since the analysis of handwriting has grown to the point where there are numerous systems and theories, it was decided to use a standardized method, Graphoanalysis.[2] In a previous investigation using this method, Galbraith and Wilson (1964) demonstrated that interrater reliabilities ranged from unacceptably low ($r = .61$) to quite high ($r = .87$). Validity data were not reported.

METHOD

Subjects

The *Ss* in this study consisted of 63 life insurance salesmen from three life insurance companies in Honolulu. All but 4 of these 63 had been selling life insurance for at least 1 yr., and all were full-time salesmen. The remaining 4 left the life insurance business after periods ranging from 2 to 12 mo.

Procedure

All salesmen were instructed to copy the following paragraph in their normal handwriting style:

> What are the various stages in the collection process? We may define the stages as early, middle, and late. In the early stages, we can assume that the debtor will pay if we merely call the obligation to his attention. We do not send a letter—just a statement with an enclosed envelope. Later we may send a second statement as a reminder. Then if no response follows, we send a letter of inquiry to find out why the debtor is delaying payment. If we receive a

[1] This study is based upon a thesis by the first author submitted to the Graduate School, University of Hawaii, in partial fulfillment of the requirements for the master's degree.

[2] Graphoanalysis is a copyrighted method of handwriting analysis by the International Graphoanalytic Society, Chicago, Illinois. Two certified Graphoanalysts, Doris Williamson and Dorothy Galbraith, performed the analyses on the handwriting samples in this study. Their cooperation is gratefully acknowledged.

reasonable explanation, we need send no further letters.

In the opinion of the two Graphoanalysts who served as judges, this paragraph provided a sufficient number of "strokes" to permit classification on selected personality traits. The total number of times that a given trait could "appear" in the test paragraph was tabulated. The *Ss* then received a numerical score for each trait on the following basis: (1) does not appear, (2) slight number of appearances, (3) moderate number of appearances, and (4) appears consistently. Scores were assigned according to the following formula:

$$X = \frac{\text{Number of judged appearances}}{\text{Total number of appearances possible}}$$

The experiment was divided into three parts. Two different approaches were tried at predicting successful salesmen, and a third attempt was made at selecting failures.

The criterion measure of job success was based upon initial year commissions on life insurance policies sold during the calendar year 1964 (first-policy-year commissions). These earnings were converted to a score based upon years of experience in selling life insurance. This was derived by having all three companies submit data indicating the amount each of their salesmen earned in first-policy-year commissions in recent years, based on years of experience for each man. These data were for all salesmen employed by the companies in the State of Hawaii, not merely *Ss* used in this study. Once this information was tabulated, a consultation with the companies yielded criterion measures based upon Table 1.

Part I

The Graphoanalysts attempted to select successful salesmen on the basis of 13 traits (see Figure 1) which they felt were necessary in the sales profession. A score referred to as a correction factor was assigned for miscellaneous undesirable traits that appeared in the writing. This score, ranging from 0 through 3, was subtracted from the grand total for the 13 traits.

Handwriting samples were coded by *E* prior to their being given to one of the analysts. Upon completion of the ratings, they were returned to *E*, recoded, then given to the second analyst for rating.

Part II

In a second attempt, the handwriting samples of the three most outstanding producers were analyzed to determine the dominant traits they judged to be present. (They were not told the results of the first analysis, nor were they given any indication that anything else would be expected of them. The analysts never saw a handwriting sample a second time without it being recoded. Periodically, samples from salesmen with unavailable criterion scores were interspersed as buffer items.) Each analyst picked out what she felt were dominant traits in the writing. A comparison of their ratings indicated that the following traits appeared consistently: *enthusiasm, like of variety, self-castigation, broad-mindedness, vanity,* and *independence.*

The remaining samples were then analyzed for these traits on the assumption that those displaying them would be successful. The analysts, working together, gave each salesman a score of 1 (trait does not appear) to 5 (outstanding on trait) for each of the above traits.

Part III

The third approach consisted of attempting to select failures. The handwriting samples of the four salesmen who had quit were analyzed by the analysts in order to discover any "inadequacies" they felt were present. The handwriting strokes indicated that these four lacked *self-confidence, tenacity, independence,* and were low in *desire for change and variety.*

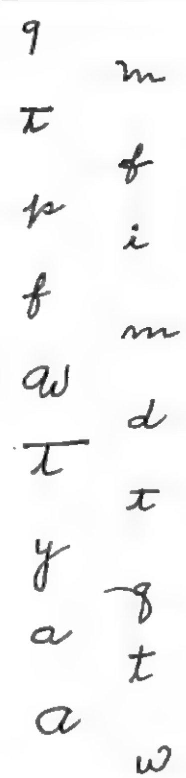
For this attempt the criterion was redefined by calling Criterion Scores 1 and 2 "apt to fail," 3 was the borderline point, and 4 and 5 criterion scores represented "successful salesmen." From the remaining handwriting samples, 23 successes, 23 failures, and 8 borderline cases were selected. The analysts, working together, were asked to deter-

TABLE 1

Criterion Scores in Terms of Earnings and Experience

Years of experience	Criterion score				
	1	2	3	4	5
1	Less than 2000	2000–3000	3000–4000	4000–5500	5500+
2	Less than 3000	3000–4000	4000–5000	5000–6500	6500+
3	Less than 3750	3750–4750	4750–5750	5750–7000	7000+
4	Less than 4500	4500–5500	5500–7000	7000–8000	8000+
5+	Less than 4500	4500–5500	5500–7000	7000–8500	8500+

Note.—Earnings given include only first-policy-year commissions for individuals exclusively selling life insurance.



Personality Trait	Determined by
Determination	Width or heaviness of downstroke
Diplomacy	Tapering off or slanting structures
Purpose	Heaviness and height of t-crossings
Persistence	Tied strokes in the form of a loop or knot
Initiative	Forward swinging stroke in a p, g, j, or y
Attention to Detail	Dot close to and directly above the letter stem
Organizational Ability	Balanced loops above and below the writing line
Analytical Ability	v-formations at the base line of writing
Sense of Responsibility	Large initial loops on capital letters
Pride	Height of d and t stems
Enthusiasm	Heaviness and length of t-crossings
Independence	Short d and t stems
Desire for Variety	Long loops below the line of writing
Self-Castigation.	Strokes whipped back to the left
Broadmindedness	Well-rounded circle formations
Vanity	Exceedingly high d and t stems
Self Confidence	Large capital letters or underlined words
Tenacity	Final hooks at the end of a word

FIG. 1. Traits and their method of determination according to Graphoanalysis.

mine if a salesman possessed or lacked each of the four traits named above.

Upon completion of the ratings, the samples were taken from the analysts and returned to them a week later when they were asked to repeat this last analysis.

RESULTS

Part I

Interjudge reliability of the analysts on each of the 13 traits was determined by a Pearson product-moment correlation. All values were significant but ranged from unacceptably low ($r = .50$) to quite high ($r = .85$).

TABLE 2

INTERJUDGE RELIABILITY ON TRAITS

Determination	.50	Analytical ability	.70
Diplomacy	.85	Sense of responsibility	.60
Purpose	.62	Pride	.50
Persistence	.65	Enthusiasm	.70
Initiative	.58	Independence	.63
Attention to detail	.63	Desire for variety	.76
Organizational ability	.58	Grand total	.51
Correction factor	.36		

Correlations of the individual traits with the criterion revealed that many of these were negative, and in no case was significance reached. These values are presented in Table 3.

TABLE 3

CORRELATIONS OF TRAITS WITH THE CRITERION

Trait	Analyst 1	Analyst 2
Determination	−.08	.19
Diplomacy	−.10	−.12
Purpose	−.15	−.10
Persistence	.02	−.15
Initiative	−.21	−.07
Attention to detail	.01	.03
Organizational ability	−.12	−.19
Analytical ability	.11	−.11
Sense of responsibility	−.18	−.13
Pride	−.03	−.05
Enthusiasm	−.07	.11
Independence	.06	.03
Desire for variety	.14	.13
Correction factor	−.18	−.06
Grand total	−.22	−.13

TABLE 4

CORRELATION OF EMPIRICALLY DETERMINED TRAITS WITH THE CRITERION

Trait	r
Enthusiasm	.08
Like of variety	.12
Self-castigation	.07
Broad-mindedness	.01
Vanity	.08
Independence	.10

The 13 traits plus the correction factor were used as predictors of the criterion scores in a multiple-regression analysis. It was decided to use the regression approach because the analysts selected the predictor traits intuitively. It was therefore assumed that the traits would have different weights in a prediction formula.

The analysis was carried out independently for each analyst. The multiple R was .42 in one case and .46 for the other. Neither was significant.

Part II

A correlation of the empirically determined traits with the criterion is presented in Table 4. A regression analysis using the six traits as predictors of the criterion yielded an R of .24. Again, this was nonsignificant.

Part III

The final phase of this study consisted of attempting to select unsuccessful salesmen or "failures." For this analysis salesmen received a tally for each of the four traits they "lacked" according to their handwriting. The handwriting samples of the borderline cases (criterion scores of 3) and the four salesmen used for determining the traits were not used in the analysis. This left a total sample of 46, half of which were "successful" and half "failures."

TABLE 5

TRAITS LACKING AMONG SUCCESSFUL AND UNSUCCESSFUL SALESMEN ACCORDING TO GRAPHOANALYSIS

Trait	"Successes"	"Failures"	χ^2
Self-confidence	11	14	.360*
Tenacity	18	30	2.520***
Independence	7	12	.842**
Desire for variety	3	10	2.769****

Note.—Since the direction of the effect was predicted, a one-tailed analysis is used. (Both trials combined.)
* $p < .28$
** $p < .19$.
*** $p < .06$.
**** $p < .05$.

The distribution of tallies is presented in Table 5. Through use of a chi-square analysis (one-tailed) it was discovered that *self-confidence* was the least discriminating trait. On that basis it was dropped from the predictive battery.

Using the three remaining traits, salesmen with two or more tallies were classified as "failures." No prediction concerning salesmen with less than two tallies was made. On this basis the analysts selected a total of 10 "failures" out of a possible 23 in their first attempt. A week later, they selected a total of 12 out of 23 when the same handwriting samples were recoded and presented to them.

TABLE 6

RESULTS OF REJECTING SALESMEN WITH TWO LOW SCORES

Trial	Correct	Incorrect	χ^2
1	8	2	2.500**
2	9	3	2.083*
1 and 2 combined	17	5	5.500***

Note.—p given for one-tailed test of significance. Nine of the 10 choices in Trial 1 were reselected in Trial 2 indicating high reliability, but casting some doubt on the appropriateness of the pooled chi-square analysis. However, familiarity effects were minimized due to the 1-wk. time lapse between trials and the total number of ratings involved.
* $p < .09$.
** $p < .06$.
*** $p < .01$.

As can be seen in Table 6, less than 50% of the total possible failures were detected. However, of those who were identified as failures, a significant number were actual failures.

DISCUSSION

The results of this study must be interpreted in light of certain assumptions. Foremost of these is that construct validity of the traits was not and probably could not be determined adequately. Rather, attempts were made at selection according to a con-

current validity model (see Cronbach, 1960). Therefore, what are referred to as "traits" are merely characteristic handwriting "strokes."

It must also be understood that the handwriting samples used in this study came from salesmen currently selling life insurance (with the exception of four). Therefore, one could assume that they met existing company standards in order to be hired in the first place. Although many failures in selling life insurance are to be expected, this study essentially consisted of testing the ability of Graphoanalysts to identify the failures who were not screened out in the selection process by the life insurance companies.

An examination of the results indicates that both attempts to select successful life insurance salesmen (Parts I and II) were unsuccessful. The first attempt used traits selected a priori as indicative of success by the Graphoanalysts, while the second attempt used traits that the analysts found were prominent in the handwriting samples of successful salesmen.

In the attempt to select failures (Part III) the results indicated that it may be possible to do this, although about half of the actual failures, as judged by the criterion, remained undetected by using the three traits mentioned herein.

It is possible that the results of this last attempt might in some way be contaminated due to repeated exposure of the analysts to the same handwriting samples, even though every possible attempt was made to minimize this. Precautionary measures included recoding of the handwriting samples prior to each analysis, inserting of "dummy" samples, and also withholding results of previous analyses from the analysts. Furthermore, no information concerning the handwriting samples or their source was made available to them.

Perhaps more important than the results obtained are the implications provided by this validity study. A compelling rationale for the results can be offered. If one assumes that the handwriting strokes are valid indications of certain indicated traits, can one necessarily assume to predict success in any field? The answer is, of course, no. Success depends not on the availability of desirable traits, but rather, on the successful application of these traits.

On the other hand, if a person lacks traits deemed essential in a given occupation, his chances of success diminish markedly. On this basis it therefore becomes possible to predict failure, but not success. Using the same rationale, no predictive statement can be made concerning persons with desirable traits, for they may succeed or fail, depending upon their utilization of these traits.

In conclusion, one might state that it *may* not be possible to predict success by using handwriting analysis, possibly due to the fault of the method or possibly due to lack of understanding of the personality types that accompany success in different occupations. It may be possible to predict failure, however, and future validity attempts might concentrate on this aspect.

REFERENCES

BELLOWS, R. *Psychology of personnel in business and industry.* Englewood Cliffs, N. J.: Prentice-Hall, 1961.

CRONBACH, L. J. *Essentials of psychological testing.* New York: Harper & Row, 1960.

FLUCKIGER, F. A., TRIPP, C. A., & WEINBERG, G. H. A review of experimental research in graphology, 1933–1960. *Perceptual & Motor Skills,* 1961, 12, 67–90.

GALBRAITH, D., & WILSON, W. Reliability of the graphoanalytic approach to handwriting analysis. *Perceptual & Motor Skills,* 1964, **19**, 615–618.

LEWINSON, T. A. The use of handwriting analysis as a psychodiagnostic technique. *Journal of Projective Techniques,* 1961, **25**, 315–329.

SONNEMANN, U., & KERNAN, J. Handwriting analysis —a valid selection tool? *Personnel,* 1962, 39, 8–14.

(Received July 5, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 3, 300-304

AUDITORY S-R COMPATIBILITY:

THE EFFECT OF AN IRRELEVANT CUE ON INFORMATION PROCESSING

J. RICHARD SIMON AND ALAN P. RUDELL

University of Iowa

2 experiments demonstrated the existence of a strong population stereotype which affected the processing of verbal commands. In a choice RT task, Ss pressed the right- or left-hand key in response to the words "right" or "left" which were presented to the right or left ear. RT was significantly faster when the content of the command corresponded to the ear stimulated than when it did not; i.e., information processing was affected by a cue irrelevant to the task itself, the ear in which the command was heard. Removing S's uncertainty regarding the ear to be stimulated resulted in significantly faster RT, and reduced but did not eliminate the effect of the irrelevant directional cue.

A fundamental problem in man-machine system design is that of integrating display and control elements so that the information presented can be processed efficiently into appropriate control action. To facilitate information processing, the human engineer seeks to conform to population stereotypes; that is, he designs displays so that they give a "natural" indication of the correct response. This occurs, for example, when the spatial arrangement of the display corresponds with that of the control (Fitts & Seeger, 1953; Garvey & Mitnick, 1955). Fitts and Seeger use the term S-R compatibility in referring to the extent to which the ensemble of stimulus and response combinations comprising the task results in a high rate of information transfer.

Up to now, the concepts of stereotype and compatibility have been applied almost exclusively in the realm of visual displays (Fitts, 1951; Loveless, 1962; McCormick, 1964). A notable exception is a study by Mudd (1963) which demonstrates that dimensions of pure tone (frequency, intensity, and interaural intensity difference) have a natural tendency to be associated with horizontal and vertical dimensions of space. Mudd's research has been concerned with reducing visual search time by providing *S* with additional relevant auditory cues (Mudd, 1965; Mudd & McCormick, 1960). The concern of the present study is with auditory S-R compatibility, that is, the role of irrelevant directional cues imbedded in the auditory display itself which interfere with information processing. Data to be presented indicate that the interpretation of a verbal directional command is affected by the ear in which the command is heard.

The major findings of the present study provide an excellent example of serendipity. The original purpose of this study was to investigate the interaction between ear stimulated and handedness in an auditory reaction-time (RT) task. A number of recent studies indicate that the right ear is more efficient than the left in perceiving verbal materials under dichotic listening conditions, that is, simultaneous presentation of different stimuli to both ears (Broadbent & Gregory, 1964; Bryden, 1963; Kimura, 1961b, 1963). This greater right-ear effectiveness has not been interpreted to indicate right-ear superiority *as such* but rather the superiority of the ear opposite the dominant hemisphere for speech. Kimura (1961a) found that, when speech was represented in the left hemisphere, the right ear was more efficient, but, when speech was represented in the right hemisphere, the left ear was more efficient. The aim of the present study was to extend these dichotic listening results to a monaural RT task. It was predicted that right-handed *S*s (presumed to be almost exclusively left-hemisphere dominant) would respond faster to a verbal command in their right ear than in their left, whereas left-handed *S*s (a majority presumed to be right-hemisphere dominant) would tend to respond faster to commands in the left ear.

EXPERIMENT I

Apparatus. The apparatus provided a measure of choice RT to a series of monaural verbal commands which were presented to *S* through Grason-Stadler TDH-39 matched earphones. Two telegraph keys were positioned 10 in. apart on a table surface in front of *S*. Instructions were to depress the right key as quickly as possible after hearing the word "right" and to depress the left key as quickly as possible after hearing the word "left." The *S* rested his forearms on the table surface and, upon the onset of a ready signal, placed his index fingers on the keys. The ready signal was presented 2 sec. prior to each command and consisted of the simultaneous onset of a warning light and a 1000-cps, 80-db. binaural tone of 500-msec. duration. A Hunter KlocKounter started when the command was presented and stopped when *S* pressed the key. Depressing the key also signaled *E* as to which key had been pressed. The instructions and verbal commands were recorded on tape and presented at 70 db. by a Sony TC 500 Stereocorder. There was a 6-sec. interval between commands.

Procedure and experimental design. Each *S* responded to the same series of 132 prerecorded commands, each command consisting of the word "right" or the word "left." The commands were presented to either the right ear or the left ear in a predetermined random sequence. The *S* had no way of knowing, prior to the presentation of each command, which ear would be stimulated or what the command would be. The first eight trials were practice trials in which the "right" and "left" commands were presented twice to each ear in a random order. For half of the *Ss* in each Sex × Handedness subgroup the earphones were reversed in order to balance out any differences which may have existed between the two stimulus channels.

Subjects. Two groups were used, one of 48 self-classified strongly right-handed *Ss* and the other of 32 self-classified strongly left-handed *Ss*. The *Ss* ranged in age from 18 to 28 (average age 19.7). There was an equal number of males and females in each handedness group. The *Ss* all passed a standard audiometric screening test in which pure tones of 500, 1000, and 2000 cps were presented to each ear separately. No *S* had a hearing loss greater than 10 db. at any one of the three frequencies tested. Each *S* also responded to a short handedness-questionnaire indicating the hand he used to write, turn a screwdriver, throw a ball, hold a toothbrush, and swing a tennis racket. All of the right-handed *Ss* reported that they used their right hand exclusively for these activities while three left-handed *Ss* reported using their right hand for one of the five activities queried.

Results. Median RTs were computed for each *S* for each of the four experimental conditions, that is, "right" and "left" commands in the right and in the left ear. Trials on which anticipations or incorrect responses occurred (about 1% of the trials) were omitted.

Table 1 summarizes the mean RT under each treatment condition for each Sex × Handedness subgroup. Analysis of variance revealed no differences as a function of the main effects of ear stimulated, command, or handedness group. There was, however, a significant difference between the sex groups—$F(1/76) = 5.02$; $p < .05$—with males responding faster than females (391 versus 426 msec.).

The predicted Ear Stimulated × Handedness interaction was not significant—$F(1/76) = 1.95$, although, as Table 1 indicates, the differences were in

TABLE 1

REACTION TIME TO VERBAL COMMANDS AS A FUNCTION OF EAR STIMULATED, HANDEDNESS, AND SEX

Group	Ear stimulated	Males			Females			*M* ear stimulated
		Command			Command			
		"Right"	"Left"	*M*	"Right"	"Left"	*M*	
Right handers	Right ear	363	413	388	400	441	421	404
	Left ear	411	375	393	445	406	426	409
	M	387	394	391	423	424	423	407
Left handers	Right ear	377	406	391	426	434	430	411
	Left ear	427	352	390	458	400	429	409
	M	402	379	390	442	417	429	410
M of handedness groups combined		393	388	391	430	421	426	

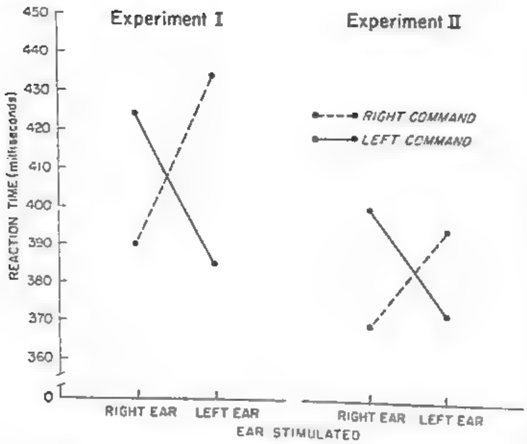


FIG. 1. Effect of ear stimulated on RT to verbal directional commands.

the expected direction with right-handers tending to respond faster to right-ear stimulation and left-handers showing the opposite tendency. A chi-square analysis indicated that a significant proportion of the right-handed group (32/48) responded faster on the right-ear trials than on the left-ear trials ($\chi^2 = 4.68$; $p < .05$). There was no tendency, however, among left-handed *Ss* to favor the left ear (15 were faster on the left-ear trials, 15 were faster on the right, and there were 2 ties), so the handedness groups did not differ significantly in the relative proportion of *Ss* showing right-ear superiority ($\chi^2 = 2.34$; $p > .10$). The analysis of variance revealed a significant Command × Handedness interaction—$F(1/76) = 11.82$; $p < .01$—which reflects the fact that right-handers tended to respond faster to the "right" command (i.e., with their right hand), and left-handers tended to respond faster to the "left" command (i.e., with their left hand).

By far the most conspicuous result was the significant Command × Ear Stimulated interaction—$F(1/76) = 226.50$; $p < .001$. The left half of Figure 1 pictures this effect. It will be noted that RT was markedly faster when the "right" command was heard in the right ear than when it was heard in the left ear (390 versus 434 msec.). Similarly, RT to the "left" command was faster when it was heard in the left ear than when it was heard in the right ear (385 versus 424 msec.)

EXPERIMENT II

The purpose of this experiment was to determine whether the irrelevant cue noted in Experiment I (i.e., ear stimulated) would continue to affect RT if *S* knew which ear would be stimulated prior to presentation of the command. It was conjectured that, as a result of removing *S*'s uncertainty as to the source of the command, the variance contributed by this irrelevant cue would be eliminated and the hypothesized interaction between ear stimulated and handedness might then be detected.

Apparatus, procedure, and experimental design. The apparatus was the same as that used in Experiment I. The procedure, however, differed in that trials were presented in blocks rather than in the random ear order employed previously. Each *S* performed on two blocks of trials, one block in which the commands were presented to his right ear and the other block in which the commands were presented to his left ear. Instructions prior to each block of trials stressed the fact that *S* would hear the commands in his right ear *only* or in his left ear *only* as the case may be. Each block consisted of 62 test trials appearing at 5-sec. intervals in which the word "right" and the word "left' appeared an equal number of times in a predetermined random sequence. Each block was preceded by four practice trials consisting of two "right" commands and two "left" commands. The instructions and verbal commands were prerecorded and presented at 70 db.

The design was an extension of a Type VI (Lindquist, 1953) with two between- and two within-*S* dimensions. Half of the 16 *Ss* in each Sex × Handedness subgroup performed the right-ear trials first while the other half performed the left-ear trials first. By simply reversing the headset, the same random series of commands was used for both right- and left-ear blocks of trials.

Subjects. The *Ss* were 64 University of Iowa undergraduate volunteers between the ages of 18 and 29 (average age 20.0). Two groups were used; one consisted of 32 strongly right-handed *Ss* and the other of 32 strongly left-handed *Ss*. There were 16 men and 16 women in each group.

All *Ss* were given a standard audiometric screening test, and none had a hearing loss greater than 10 db. at any one of three "speech" frequencies: 500, 1000, and 2000 cps. The *Ss* then completed the same handedness-questionnaire that was used in Experiment I. All of the right-handed *Ss* and 26 left-handed *Ss* reported using their preferred hand exclusively for the 5 activities queried; whereas 6 left-handed *Ss* reported using their right hand or either hand for 1 or 2 of the tasks.

Results. Median RTs were computed for each *S* under the four experimental conditions, that is, "right" and "left" monaural commands in the right and left ear. Analysis of variance provided no evidence of the predicted Ear Stimulated × Handedness interaction. There were also no differences due to the main effects of ear stimulated, command, or handedness group. The Command × Handedness interaction, which was significant in Experiment I, was not replicated. Males tended to respond faster than females (372 versus 396 msec.), but the difference failed to reach an acceptable level of significance—$F(1/60) = 2.64$; $p > .10$. The difference between right- and left-ear trials was different for males than for females—$F(1/60) = 5.98$; $p < .05$.

As in Experiment I, the major source of variance was the Command × Ear Stimulated interaction—$F(1/60) = 83.46$; $p < .001$—which is pictured in the right half of Figure 1. It will be noted that the compatible reactions, that is, "right" in the right ear and "left" in the left ear, were significantly faster than the incompatible reactions, that is, "right" in the left ear and "left" in the right ear. Mean RT for the two compatible conditions was 371 msec. versus 397 msec. for the two incompatible conditions.

Comparison of Experiments I and II. Figure 1 facilitates a comparison of the results of the two experiments. To reiterate, the major difference in procedure between the two was that, in the latter experiment, instructions were used to remove *S*'s uncertainty regarding the ear in which the command would appear. In the former experiment, on the other hand, *S* did not know prior to hearing the command which ear would be stimulated. An overall analysis of the combined data from the two experiments indicated that *Ss* responded significantly faster in Experiment II than in Experiment I (384 versus 408 msec.)—$F(1/140) = 5.07$; $p < .05$. Figure 1 indicates that while both compatible and incompatible reactions were faster under the conditions of Experiment II, there was a relatively greater speedup of the incompatible reactions. This latter effect was reflected in a significant Command × Ear Stimulated × Experiment interaction—$F(1/140) = 14.97$; $p < .001$. In the random ear order presentation employed in Experiment I, the average difference between compatible and incompatible reactions was 42 msec. This difference was reduced to 26 msec. in Experiment II where *Ss* were told in advance which ear would be stimulated.

DISCUSSION

Results clearly indicated that speed of processing verbal commands (i.e., the words "right" and "left") was affected by a cue irrelevant to the task itself, that is, the ear in which the command was heard. RT was significantly faster when the content of the command corresponded to the ear stimulated (i.e., "right" in the right ear or "left" in left ear) than when it did not (i.e., "right" in left ear and "left" in right ear). These results suggest the existence of a strong natural tendency to associate right-ear stimulation with a right-hand response and left-ear stimulation with a left-hand response. When the verbal command called for a response which corresponded to this population stereotype, *Ss* responded rapidly. However, when conflict existed between the symbolic cue (content of command) and the irrelevant directional cue (ear stimulated), a marked delay in responding occurred. It was as if *Ss* had to inhibit the stereotypic directional response before reacting to the symbolic content of the command.

The data underline the potency of the stereotype in question. The irrelevant cue was particularly influential in Experiment I where *Ss* were uncertain as to which ear would be stimulated. Removing this uncertainty in Experiment II significantly reduced but by no means eliminated the effect.

Questions still remain as to the generality of the findings. Is the stereotype specific to a bimanual task such as the one employed in the present study where *S* responds with the body member on the same or the opposite side as the ear stimulated? Or would the same irrelevant cue operate in a unimanual task where, for example, *S* moves a lever to the right or left from the midline of his body in response to verbal commands?

Auditory displays have not been widely used for providing directional information in man-machine systems (McCormick, 1964). "Flybar" (Chapanis, Garner, & Morgan, 1947, ch. 9) represented an early attempt to provide pilots with spatial information by means of complex auditory signals. Present results suggest that additional basic studies are urgently needed. One potentially promising avenue for research would involve attempting to increase the speed of processing directional information from an auditory display by *utilizing* the cue provided by the ear stimulated. One wonders whether *Ss* would respond faster to a combination of directional and symbolic cues ("right" command in the right ear or "left" in the left ear) than to directional or symbolic cues alone.

Other results bear mentioning because of their relation to previous research. The finding that men responded significantly faster than women supports the oft-noted tendency for male superiority on RT tasks (Teichner, 1954; Woodworth & Schlosberg, 1955). The tendency for the right-handed group to respond faster to right-ear stimulation is consistent with recent findings of ear preference in auditory perception (Borkowski, Spreen, & Stutz, 1965; Bryden, 1963; Dirks, 1964) and supports the view that *Ss* tend to tune in with their right ear when they are uncertain as to which ear will be stimulated.

REFERENCES

Borkowski, J. G., Spreen, O., & Stutz, J. Z. Ear preference and abstractness in dichotic listening. *Psychonomic Science,* 1965, **3**, 547–548.

Broadbent, D. E., & Gregory, M. Accuracy of recognition for speech presented to the right and left ears. *Quarterly Journal of Experimental Psychology,* 1964, **16**, 359–360.

Bryden, M. P. Ear preference in auditory perception. *Journal of Experimental Psychology,* 1963, **65**, 103–105.

Chapanis, A., Garner, W. R., & Morgan, C. T. *Applied experimental psychology.* New York: Wiley, 1949.

Dirks, D. Perception of dichotic and monaural verbal material and cerebral dominance for speech. *Acta Oto-Laryngologica, Stockholm,* 1964, **58**, 73–80.

Fitts, P. M. Engineering psychology and equipment design. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951. Pp. 1287–1340.

Fitts, P. M., & Seeger, C. M. S-R compatibility: Spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology,* 1953, **46**, 199–210.

Garvey, W. D., & Mitnick, L. L. Effect of additional spatial references on display-control relationships. *Journal of Experimental Psychology,* 1955, **50**, 276–282.

Kimura, D. Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology,* 1961, **15**, 166–171. (a)

Kimura, D. Some effects of temporal-lobe damage on auditory perception. *Canadian Journal of Psychology,* 1961, **15**, 156–165. (b)

Kimura, D. A note on cerebral dominance in hearing. *Acta Oto-Laryngologica, Stockholm,* 1963, **56**, 617–618.

Lindquist, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

Loveless, N. E. Direction-of-motion stereotypes: A review. *Ergonomics,* 1962, **5**, 357–383.

McCormick, E. J. *Human factors engineering.* (2nd ed.) New York: McGraw-Hill, 1964.

Mudd, S. A. Spatial stereotypes of four dimensions of pure tone. *Journal of Experimental Psychology,* 1963, **66**, 347–352.

Mudd, S. A. Experimental evaluation of binary pure-tone auditory displays. *Journal of Applied Psychology,* 1965, **49**, 112–121.

Mudd, S. A., & McCormick, E. J. The use of auditory cues in a visual search task. *Journal of Applied Psychology,* 1960, **44**, 184–188.

Teichner, W. H. Recent studies of simple reaction time. *Psychological Bulletin,* 1954, **51**, 128–149.

Woodworth, R., & Schlosberg, H. *Experimental psychology.* New York: Holt, 1955.

(Received August 8, 1966)

Manuscript Accepted for Publication in the *Journal of Applied Psychology*

Accuracy of Manual Entries in Data-Collection Devices: William A. Smith, Jr.*: Director, Computing Laboratory, Lehigh University, Bethlehem, Pennsylvania 18015.

Searching for Newspaper Headlines Printed in Capitals or Lower-Case Letters: E. C. Poulton*: Assistant Director, Medical Research Council, Applied Psychology Research Unit, 15 Chaucer Road, Cambridge, England.

Some Effects on Business Gaming of Previous Quasi-T Group Affiliations: Samuel D. Deep, Bernard M. Bass,* and James A. Vaughan: Graduate School of Business, University of Pittsburgh, 401 Bruce Hall, Pittsburgh, Pennsylvania 15213.

Vocational Interests of Female Computer Programmers: Dallis K. Perry* and William M. Cannon: Manager, Personnel Research, System Development Corporation, 2500 Colorado Avenue, Santa Monica, California 90406.

* Asterisk indicates author for whom the address is supplied.

journal of applied psychology

VOL. 51, No. 4, Part 1 of 2 Parts AUGUST 1967

MULTIDIMENSIONAL SCALING ANALYSIS OF DECISION STRATEGIES IN THREAT EVALUATION[1]

JOSEPH W. RIGNEY AND CHARLES H. DEBOW

University of Southern California, Los Angeles

Similarities analysis was used to identify the combinatorial strategy used by naval officers in evaluating the relative threat of paired air-raids presented on an air summary plot. 63 Combat Information Center (CIC) watch officers judged the similarity, in terms of threat value, of all different pairs of 20 air raids, varying in range, course, bearing, composition, altitude, and speed. The threat value of each raid subsequently was estimated, using a 9-point scale. It was found that the judgments of threat value were based primarily on range and course of the raids.

For decision tasks in which the outcome is of great consequence and the time for deciding is short, as in many military, business, and medical decisions, the strategy the decision maker uses is of special interest. Given a number of variables in the decision situation, does he consider all of them or does he simplify the situation, when he has to form a weighted combination of variables in order to arrive at a decision? Shepard (1964b) remarks on the "obvious disparity between the effortlessness and surety of most perceptual decisions and the painful hesitation and doubt characteristic of these 'higher level' decisions [p. 263]."

The present study was designed to analyze the judgments of military decision-makers in a context highly analogous to the actual decision-making environment. The study was concerned with threat evaluation and weapons assignment in anti-air warfare. In the study, similarities analysis, or multidimensional scaling (Kruskal, 1964; Shepard, 1962a, 1962b; Torgerson, 1958, 1965), was used to analyze how decision makers go about assigning weights to a number of attributes of air raids and combining these weights into a single measure of the relative threat of these raids.

There is a great deal of evidence from studies of similar judgmental problems in other contexts that humans unknowingly simplify the decision situation by concentrating on one or two salient characteristics of the objects and ignoring the others (DeSoto, 1961; Meehl, 1954; Miller, 1956; Shepard, 1964b; Yntema & Torgerson, 1961).

Cliff (1965) investigated a similar problem using a different kind of stimuli. He determined a multidimensional space from dissimilarity judgments of pairs of facial expressions and also had ratings made of the intensity of the emotions conveyed in each expression. He found that these latter, independently obtained, unidimensional judgments could be reproduced from the locations of the expressions in the multidimensional space.

In the context of this study, this simplification might not necessarily be bad from an operational standpoint, if the one or two variables that were selected were indeed the most important ones in determining threat, and if a simple combination of these weights could predict most of the threat variance. But it is still highly desirable to know if this is the strategy used, and, if so, exactly what are

[1] The results reported in this paper are a part of a research program (Rigney & DeBow, 1966) sponsored by the Personnel and Training Branch, Office of Naval Research; and the Psychological Research Branch, Bureau of Naval Personnel, under Contract Nonr-228(22).

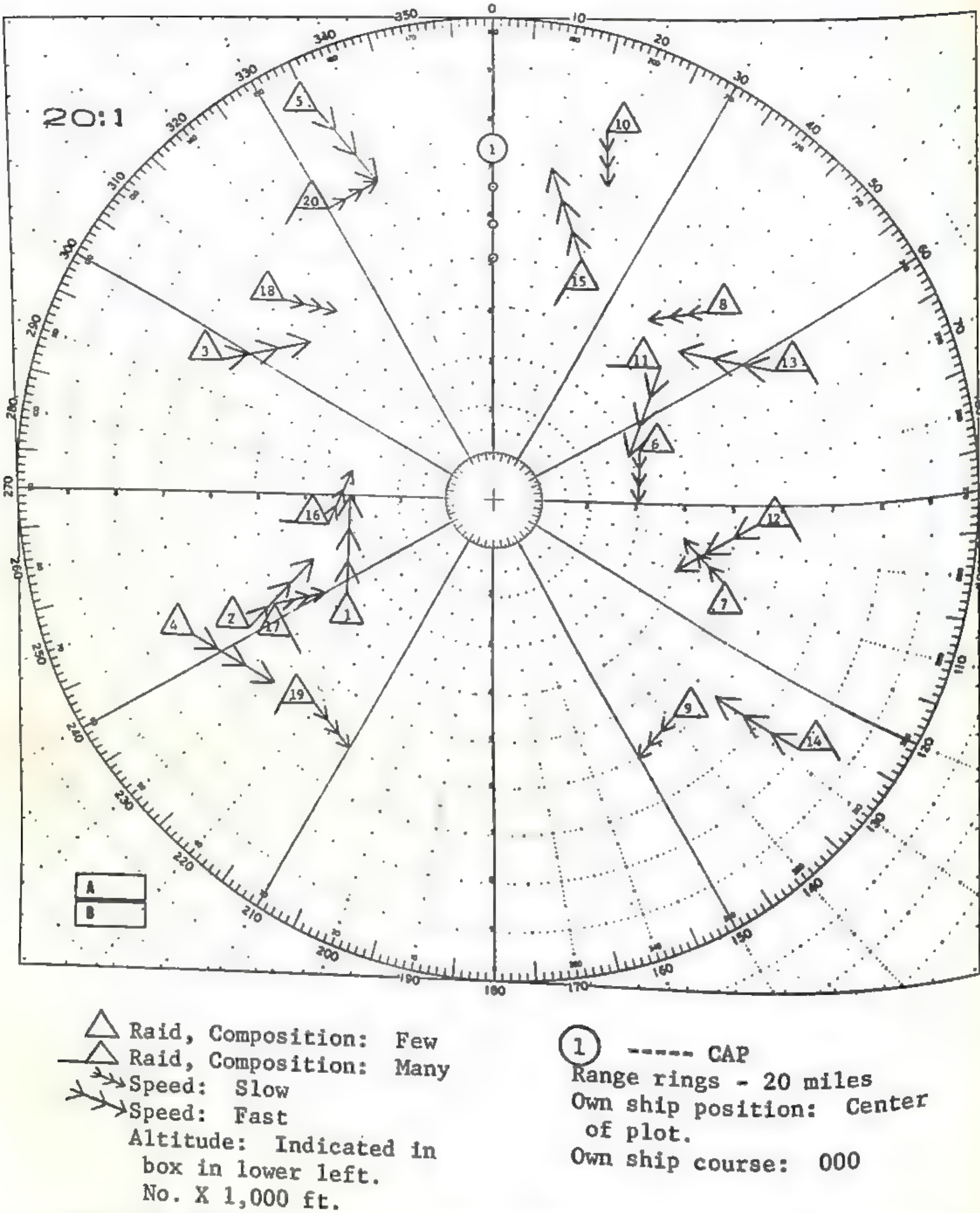


FIG. 1. Example of air summary plot. (The legend translates symbols on display. All raids are shown in this figure. During the experimental presentation, only one pair of raids was shown at a time.)

the weights and how they do combine. Other purposes of the study were to: (*a*) determine the suitability of using the multidimensional scaling technique for analyzing this aspect of decision making, and (*b*) estimate the ability of decision makers to make reliable similarity judgments of stimuli as complex as those used.

METHOD

Subjects

The subjects (*Ss*) in this study were 63 naval officers and noncommissioned officers. The 63 *Ss* were enrolled in the Combat Information Center Watch Officer Course, C-351, at the Fleet Anti-Air Warfare Training Center, San Diego, California.

TABLE 1

COMBINATIONS OF VARIABLES USED TO DESCRIBE INDIVIDUAL RAIDS

Stimulus	Composition	Speed (knots)	Bearing	Course	Altitude (× 1000 ft.)	Range (miles)
1	Few	800	270	0	30	60
2			250	40	10	80
3			310	80	30	100
4			230	120	10	120
5		800	340	140	30	140
6		400	90	180	10	60
7			100	320	30	80
8			40	260	10	100
9			150	220	30	120
10	Few	400	20	180	10	140
11	Many	800	70	200	30	60
12			110	240	10	80
13			50	280	30	100
14			130	300	10	120
15		800	10	340	30	140
16		400	280	20	10	60
17			240	80	30	80
18			320	100	10	100
19			210	140	30	120
20	Many	400	340	60	10	140

Note.—This order was randomized in all presentations.

There were four classes, separated by 2-wk. intervals. The data were collected from *S*s on the last day of each 4-wk. session.

The 4-wk. course was designed to familiarize naval officers with the equipment, operations, displays, and symbology utilized in the shipboard Combat Information Center (CIC).

Apparatus

The *S*s were seated in the same classroom in which the 4-wk. course had been administered. A 35-mm. slide projector, with an automatic interval timer set for 20 sec., was used to project the stimuli onto a screen at the front of the classroom. The stimuli were projected in a manner which maintained the same size relationships of the symbols and plotting board in the CIC. Room illumination was sufficient for *S*s to make written responses on their answer sheets and still see the projected stimuli clearly.

Each projected slide contained a pair of symbols, representing air raids, plotted on a CIC air summary plot. The slides were color photographs of an actual CIC display. The pairs of raids covered all possible combinations (190) of 20 individual raids. Each raid was replicated once as a check on reliability of *S*'s judgments. Thus, there were 210 slides, each projected for 20 sec. Figure 1 depicts each of the 20 raids employed in the study.

The characteristics of the raids varied in terms of six variables: range, bearing, course, speed, altitude, and composition (number of aircraft in the raid). The values of each of these variables are given in Table 1.

Procedure

The *S*s were required to make two types of judgments. In the first judgment, a similarity value, ranging from 1 through 9, was to be assigned to each pair of the 210 pairs of raids, presented one pair at a time. This value concerned the degree to which members of a pair were similar with respect to threat to the *S*'s ship. The value of 1 represented raids very similar in threat; 9 represented raids very different in threat. The numbers between 1 and 9 represented intermediate gradations of similarity.

Prior to the administration of the first task, *S*s reviewed a "summary of situation" which contained necessary information concerning ship position, weather, sea state, friendly aircraft, weapons available, and the general nature of potential threats.

The second judgment involved estimating the threat value of each of the raids, presented individually for 20 sec. The *S*s were asked to assign threat values ranging from 1 to 9 to each raid; 1 represented minimum threat, and 9 represented maximum threat. This was done to provide a threat reference or composite scale, for interpreting the results of the multidimensional analysis.

RESULTS

Reliability

Since the outcome of multidimensional scaling is dependent upon relatively stable similarity judgments, it was necessary to estimate the extent to which *S*s could make reliable

TABLE 2

LOADINGS OF 20 AIR RAIDS ON THE FIRST DIMENSION AND MEANS OF THREAT VALUES ASSIGNED THE RAIDS BY 32 CIC OFFICERS

Raid	Mean threat value	First dimension
1	6.19	.32
2	7.13	.36
3	6.19	−.01
4	3.56	−.47
5	4.63	−.28
6	5.97	−.19
7	6.59	.20
8	5.00	−.06
9	1.84	−.76
10	3.78	−.46
11	8.19	.58
12	7.13	.41
13	6.25	.26
14	6.84	.48
15	2.47	−.69
16	7.41	.51
17	7.06	.48
18	5.91	.27
19	2.03	−.66
20	2.31	−.65

responses to the paired stimuli. The 20 replicated items yielded stability reliabilities that, for half the sample, were below the .05 level of significance ($r = .48$). Data for this half of the sample, 31 *Ss*, were not used in the multidimensional analysis.

The principal objective of the study was to determine the extent to which the six variables—range, bearing, course, composition, speed, and altitude—influenced *S*'s judgments of raid threat value. Considering the evidence from other studies, it was assumed that the *Ss* would simplify the judgmental problem by ignoring all but the one or two most important of the six variables.

The multidimensional scaling program [2] by which the data were analyzed provides two indicators for determining minimal dimensionality. A large break in the size of latent roots suggests the rank of the matrix. And, a goodness-of-fit index indicates the degree to which a given *n*-dimensional solution fits the interstimulus distances derived from *S*'s judgments. An examination of the latent roots revealed a large difference in magnitude between the first and second roots, suggesting that a one-dimensional solution might validly represent the data. As would be expected, the goodness-of-fit index improved with the number of dimensions that were extracted, being higher for six than for one.

The loadings of the 20 air raids on the first dimension are given in Table 2, along with the threat scale obtained from the second task given the *Ss*, that of rating the threat value of each single raid. This composite scale of threat of the 20 stimuli was used as the criterion variable in a step-wise multiple-regression analysis in which the first two dimensions of the multidimensional analysis were the predictor variables.

It is obvious from inspection of Table 2 that the threat scale and the first dimension are highly correlated. This correlation is .98; the first dimension predicts 96% of the variance in the threat scale. The multiple correlation between the first two dimensions and the threat scale was still only .98. Obviously, the first dimension accounts for practically all the variance in the threat scale, and the prediction is not improved by adding another dimension.

Multiple-regression analyses were performed in order to identify the raid variables composing the first dimension (Table 3). The *Ss* considered range to be the most important determinant of threat, with course next in importance. The other four raid variables added nothing of significance to the multiple *R* with

TABLE 3

PARTIAL CORRELATIONS, MULTIPLE-REGRESSION COEFFICIENTS, AND PROPORTIONS OF R^2 PREDICTED FOR THE MULTIPLE CORRELATION OF DIMENSION ONE WITH RANGE AND COURSE VARIABLES: $R = .93$

	Partial correlation coefficients	Multiple-regression coefficients	Proportions of R^2 predicted
Range (X_1)	.90	.52	.637
Course (X_2)	.78	.31	.223

Note.—Multiple-regression equation: $Y' = .00 + .52X_1 + .31X_2$.

[2] W. S. Torgerson and G. E. Mueser. Informal notes on Torgerson & Mueser's IBM 7090 Program for Multidimensional Scaling. Mimeographed paper, 1962.

this dimension. The first dimension can be interpreted as a composite of range and course.

CONCLUSIONS

1. Multidimensional scaling can be used to identify the combinatorial strategy used by a group of CIC officers in making comparative judgments of threat value from characteristics of paired raids. In effect, the procedure recovered in the first dimension the threat scale used by the group. This dimension correlated .98 with the composite threat scale. A multiple-regression analysis of this first dimension revealed that it was a composite of range and course which accounted for 64% and 22%, respectively, of the predicted variance. The multiple R was .93. Evidently, *S*s in the study simplified the decision situation by concentrating on only two of the six raid characteristics when estimating threat value and did not utilize the other four variables to any significant extent.

2. The inability of many *S*s to make satisfactorily reliable comparative judgments of threat value of paired raids might be considered to have significant implications for training, in addition to the problem it poses for the application of multidimensional scaling to analyzing judgments based on these relatively complex stimuli. The question for training would be whether or not *S*s can be trained to make more reliable threat judgments, and if so, by what methods. The problem for future applications of similarities analysis to this type of stimulus material is to improve the reliability of judgments. Although the type of display and the symbols used in the study were those used in shipboard CICs, they possibly did not represent the raid characteristics in a fashion that was optimum for the task given the CIC officers. For example, range, course, and bearing probably were easier to visualize than altitude and composition. Since the display was not three-dimensional, altitude was not graphically presented, and the distinction between one and many aircraft depended on a very small extension of one side of a triangular symbol.

However, the primary problem for reliability may lie in the nature of the stimulus material. It has been hypothesized (Shepard, 1964a) that *S*s attend first to one subset of attributes and then to another independent or overlapping subset on successive presentations of complex stimulus material. It is possible that they change their strategies between successive presentations. Perhaps more attentional fluctuation and strategy changing is to be expected when the judgmental task concerns a conceptual rather than a perceptual structure.

In this regard, the application of similarities analysis in this study is a type of application discussed by Torgerson (1965) as more cognitive than perceptual. He discusses the problems of applying the technique to describing cognitive structures, suggesting that similarity in this case is

> sensitive to all of the delicate problems of attitude and strategy involved in decision-making tasks in general. Here, degree of similarity is not an invariant relation between a pair of stimuli, but rather depends on such things as stimulus context [pp. 389–390].

Thus, in this case, the potential threat of the raids supplied a strong context which determined the characteristics of the similarity judgments. If another context were supplied for the judgments, it is highly likely they would be different. This and the reliability question are both deserving of further study.

3. Steps were taken in this study toward objective and quantitative analysis of important human-decision processes. In addition to the implications for further research, the analytical procedures have potential uses as criterion measures in training threat evaluators, and as sources of dependent variables in the design of displays and symbols.

REFERENCES

CLIFF, N. *Multidimensional scaling and cognition: I. Relations among judgments obtained under varying conditions.* Los Angeles: University of Southern California, Report No. 1, Contract Nonr-228(32), November 1965.

DESOTO, C. B. The predilection for single orderings. *Journal of Abnormal and Social Psychology,* 1961, 62, 16–23.

KRUSKAL, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika,* 1964, 29, 115–129.

MEEHL, P. E. *Clinical versus statistical prediction.* Minneapolis: University of Minnesota Press, 1954.

MILLER, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review,* 1956, 63, 81–97.

RIGNEY, J. W., & DEBOW, C. H. *Decision strategies in AAW: I. Analysis of air threat judgments and weapons assignments.* Los Angeles: University of Southern California, Electronics Personnel Research Group, April 1966. (Tech. Rep. 47)

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika,* 1962, 27, 125–140. (a)

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika,* 1962, 27, 219–246. (b)

SHEPARD, R. N. Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology,* 1964, 1, 54–87. (a)

SHEPARD, R. N. On subjectively optimum selection among multiattribute alternatives. In G. L. Bryan & M. W. Shelley (Eds.), *Human judgments and optimality.* New York: Wiley, 1964. Pp. 257–281. (b)

TORGERSON, W. S. *Theory and methods of scaling.* New York: Wiley, 1958.

TORGERSON, W. S. Multidimensional scaling of similarity. *Psychometrika,* 1965, 30, 379–393.

YNTEMA, D. B., & TORGERSON, W. S. Man-computer cooperation in decisions requiring common sense. *IRE Transactions on Human Factors in Electronics,* March 1961, HFE-2, 20–26.

(Received July 15, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 4, 311–315

IMPROVING INSPECTOR PERFORMANCE THROUGH TRAINING AND VISUAL AIDS

FRED B. CHANEY AND KENNETH S. TEEL

Autonetics, A Division of North American Aviation, Inc., Anaheim, California

An experimental study was performed to evaluate, singly and in combination, the effectiveness of a 4-hr. training program and a set of visual aids designed to improve the performance of 27 experienced machined-parts inspectors. The criterion used was the percentage of true defects detected in a selected sample of machined parts. Findings indicated that (a) use of training alone resulted in a 32% increase in defects detected, (b) use of visual aids alone resulted in a 42% increase, and (c) use of both resulted in a 71% increase, while (d) performance of the control group did not change.

Accurate inspection performance is essential to the success of any organization which manufactures, assembles, and sells a product. It is particularly critical, however, to those organizations that build precision equipment which is expected to operate for long periods of time under a wide variety of environmental conditions. For this reason, the authors have, for the past several years, been conducting a series of studies designed to improve inspection performance.

The purpose of the study reported herein was to evaluate the effectiveness of training and visual aids in improving the performance of machined-parts inspectors.

Machined-parts inspectors are individuals whose job is to examine precision machined parts for defects which might make them unsuitable for use. The two items used in this study are shown in Figure 1; both are typical of the kinds of parts inspected.

These parts are examined for both objective and subjective defects. Included in the objective category are such defects as mislocated holes, threaded holes, lack of parallelism and concentricity, and improper dimensions. In the subjective category are such things as poor finish, scratches, and nicks. A part may be rejected for either objective or subjective defects; experience indicates, however, that objective defects are far more serious and therefore are the reasons for most rejections. A rejected part is either returned to manufacturing for rework or scrapped. Although rework and scrappage are expensive even at this point, it is far more expensive to allow a defective part to be incorporated into a delivered product and subsequently to have that product returned by the customer. Thus, while the stated goal of inspection is to maximize the number of legitimate defects detected and to minimize the number of false ones, it appears to be better, when in doubt, to err on the side of rejecting an item rather than accepting it.

One of the reasons machined-parts inspectors were selected for study was that previous informal investigations had indicated that these inspectors were identifying less than 50% of the objective defects in the items reviewed.

METHOD

Job-sample performance measures were used to divide 27 machined-parts inspectors into four matched groups. One group served as a control. A second was given a specially designed 4-hr. training program. A third was given both the training and visual aids. Subsequently, equivalent job-sample measures were again obtained for each of the groups. The Control group used standard drawings and specifications as inspection aids for both the "before" and "after" measurements. The training and/or visual aids were introduced in the other three groups as part of a continuing research program which had been carried out, with full management support, for 2 yr. prior to the study. Consequently, *Ss* accepted the experimental treatments as part of their normal work environment, not as implied criticism of their previous performance. All *Ss* carried out their normal duties during the 6 mo. between testings. Statistical comparisons were made of the differences in performance among groups, and of the differences in "before-after" performance within each of the groups.

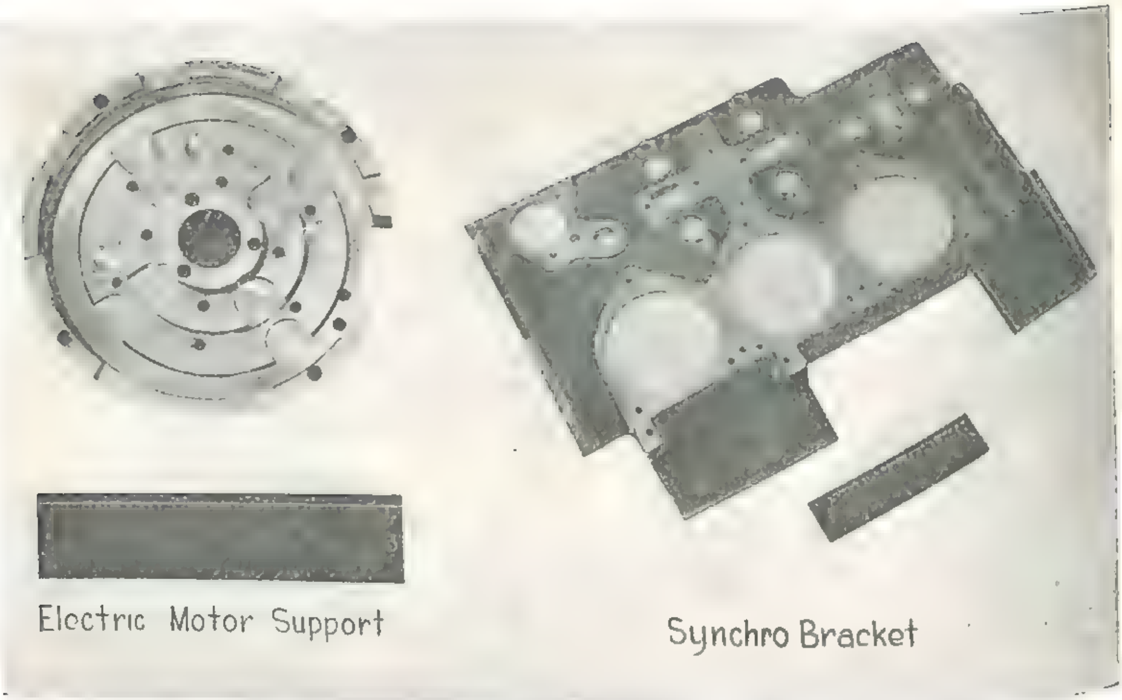


FIG. 1. Sample parts.

Subjects

The *Ss* were 27 machined-parts inspectors, the full population available in the organization. Their median age was 49, with a range from 28 to 62. Their median inspection experience was 8 yr., with a range from 3 to 20. Their median tenure with the company was 6 yr., with a range of 4–17. Thus, *Ss* were mature, experienced inspectors whose performance presumably had leveled out. Therefore, it seems safe to assume that incidental learning during the course of the study probably had little or no effect on the final performance levels of the various groups.

Job-Sample Test

Four machined parts were fabricated for use in the study. Two were brackets; the other two were motor supports. One of each was shown in Figure 1. All four items were fabricated to contain a representative sample of known defects. Each of the items was then inspected independently by four men, two leadmen and two engineering supervisors. The four then met as a group and agreed on master lists of characteristics to be inspected and defects present in each of the four parts. Each of the four parts had approximately 100 characteristics to be inspected and 34 defects.

Both the "before" and "after" measures for each of the groups were obtained by having each *S* independently inspect one bracket and one support. The initial measures were obtained with half of the *Ss* inspecting Support 1 (S_1) and Bracket 1 (B_1), and the other half inspecting Support 2 (S_2) and Bracket 2 (B_2). The final criterion measures were obtained by having *Ss* inspect the alternate parts they had not inspected previously; thus, those who inspected S_1 and B_1 initially, inspected S_2 and B_2 later, and vice versa.

Group Matching

The initial job-sample test yielded four measures of performance for each of the 27 *Ss*. The measures were (*a*) percentage of objective defects detected, (*b*) percentage of subjective defects detected, (*c*) number of false detections, and (*d*) inspection time. Since primary interest was in percentage of objective defects detected, *Ss* were initially sorted into four equivalent groups on the basis of that criterion alone. Further matching was unnecessary because statistical tests revealed that the groups did not differ significantly on any of the other three measures. The equivalence of the four groups is illustrated in Table 1.

Training Program

The training program consisted of four 1-hr. sessions conducted by the inspection supervisor. Each session included a lecture, demonstrations, and a question-and-answer period. The four sessions dealt, respectively, with the topics of (*a*) precision measurement, (*b*) thread gaging, (*c*) use of the Sheffield Internal Comparator, and (*d*) interpretation of drawing notes. The sessions were conducted once a week for 4 wk. Approximately a month elapsed between the end of the training and the administration of the final job-sample test.

TABLE 1

GROUP PERFORMANCE LEVELS BEFORE THE STUDY

Performance measure	Group means				Analysis of variance		
	Control ($N = 7$)	Training ($N = 6$)	Visual Aids ($N = 7$)	Training plus Visual Aids ($N = 7$)	*MS* groups	*MS* error	F^a ratio
Percentage of objective detections	35.7	30.5	37.6	34.3	57.3	97.6	.59
Percentage of subjective detections	53.6	51.5	47.0	54.0	77.1	75.0	1.03
No. false detections	2.3	7.0	5.0	2.7	30.9	14.5	2.13
Inspection time (hours)	11.4	12.7	12.7	11.1	4.8	8.7	.55

[a] $F_{.90}(3/23) = 2.34$.

The course contained no information on the specific defects contained in the job-sample parts, and none of the *Ss* saw those parts except during the test sessions.

Visual Aids

The visual aids consisted of a series of simple drawings of the sample parts—six for the bracket and eight for the support. One of those drawings is shown in Figure 2.

The dimensions and tolerances for each characteristic to be inspected were placed on the drawings to minimize the need for calculation or reference to other materials; these data typically had been presented on one master blueprint and in written instructions. Furthermore, similar items were grouped on a single page to encourage the inspector to examine the part fully for one class of defects before reviewing it for others. The visual aids were prepared by an experienced inspector who was not an

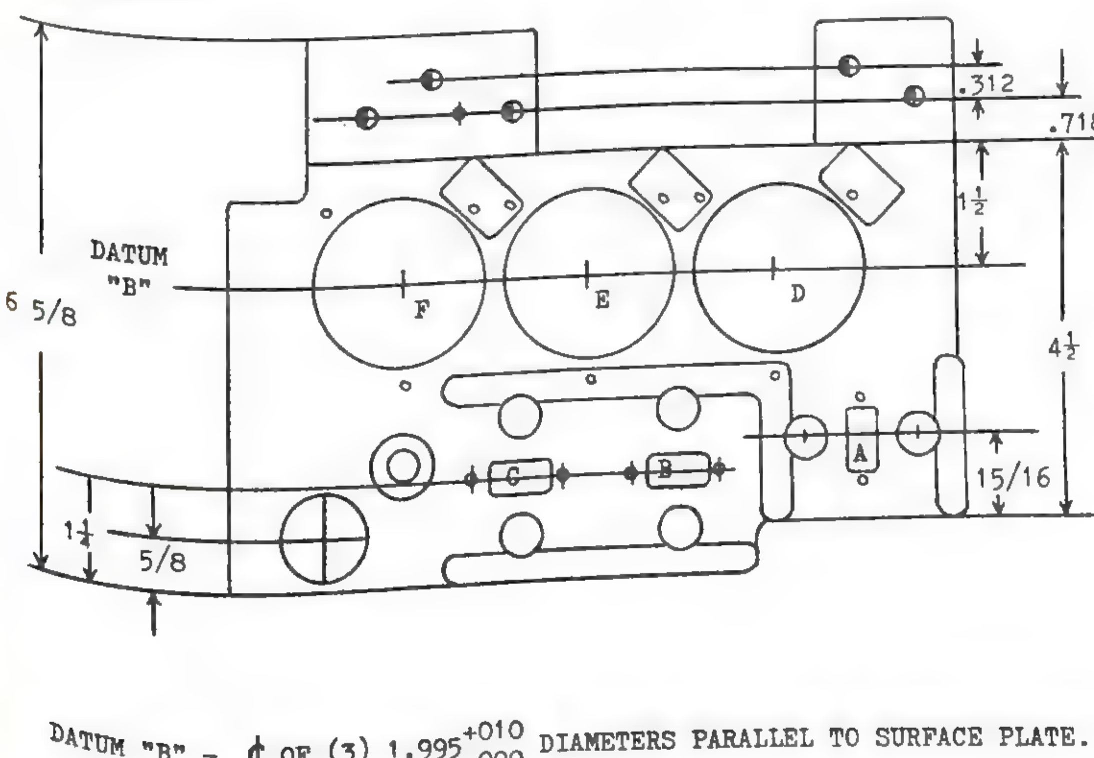


FIG. 2. Sample inspection drawing.

TABLE 2

OBJECTIVE DEFECT DETECTION

Group	Percentage of defects detected		Percentage of change
	Before	After	
Control	35.7	33.4	−6
Training	30.5	40.3	32
Visual Aids	37.7	53.4	42
Training plus Visual Aids	34.3	58.6	71

S in the study and who did not know what specific defects were contained in the sample parts.

No training was given in the interpretation or use of the visual aids. The *Ss* were simply given the drawings immediately before their final inspection of the two parts and told to use them in carrying out the inspection.

RESULTS

Since four different measures of performance were obtained, the results are presented separately for each.

Objective Defects

Results indicated clearly that use of either training or visual aids resulted in significant gains in objective detection performance, and that use of both resulted in even greater gains. These findings are presented in Table 2. It will be noted that there was no significant change in the performance of the Control group, but that there were increases of 32, 42, and 71%, respectively, in the detection of objective defects in the Training, Visual-Aids, and Training-plus-Visual-Aids groups. The *F* ratio of 10.2 for differences among the four groups was significant beyond the .001 level. "Before" and "after" comparisons within groups revealed that the increase in the Training group was significant at the .05 level, and those in the Visual-Aids and Training-plus-Visual-Aids groups were significant at the .01 level. Thus, the experimental treatments clearly resulted in significant gains in detection of objective defects.

Analysis of difference scores revealed that improvement in the performance of the Training group did not differ significantly from that of the Visual-Aids group. It also revealed that improvement of the Training-plus-Visual-Aids group was significantly greater ($p < .05$) than that of the Training group, but not significantly different from that of the Visual-Aids group. Thus, the data suggest that the visual aids may have contributed more heavily to the improved performance of the Combination group than did the training.

It should be noted that the initial performance levels of the four groups in this study were similar to those noted in previous studies of machined-parts inspection. Lawshe and Tiffin (1945) reported accuracies of 9–66% for 20 different measurement techniques. These and other findings strongly suggest that stated accuracies for many precision inspection techniques have been established under "ideal" conditions, rather than under those of the typical working environment. Thus, while the initial levels reported in this study may appear low, they probably are representative of those achieved by machined-parts inspectors under normal working conditions. The final levels, on the other hand, reveal how significant improvements in performance can be effected at relatively little cost.

Subjective Detections

No significant changes occurred in subjective defect detection performance. Changes ranged from a 3% decrease to a 7% increase; the overall level remained at approximately 50% detection of subjective defects. The *F* ratio of 1.02 for differences among the groups was not significant.

This finding is not surprising, because both the training program and the visual aids were designed solely to improve detection of objective defects. Previous studies conducted at Autonetics have indicated, however, that subjective defect detection performance can be significantly improved by providing inspectors with photographs of items containing such defects; no photographs were furnished in this study.

False Detections

No significant differences were noted among the four groups in numbers of false defects reported. Changes ranged from a 1% decrease

to an 8% increase; the F ratio of 2.19 failed to achieve significance. Thus, the gains in objective detection performance apparently were obtained without a corresponding increase in false detections.

Inspection Time

No significant differences were obtained in inspection times. The times ranged from 10.4 to 15.3 hr.; however, the F ratio of 2.32 failed to achieve significance ($F_{.05} = 3.03$). The gains in objective detection performance, therefore, were obtained without significant increase in inspection time and cost.

CONCLUSIONS

These findings demonstrate clearly that inspection performance can be improved by giving inspectors a short training program designed to correct deficiencies identified through use of objective performance measures. They also show that gains can be obtained by providing inspectors with simplified drawings which show clearly the characteristics to be inspected and the tolerances for each. Furthermore, they reveal that even greater gains can be effected by utilizing both the training and the visual aids.

Finally, they highlight the fact that industrial/experimental psychologists can make significant contributions to the solution of inspection problems. Furthermore, the benefits to the company can be measured in dollar terms; in a 2-yr. period an investment of approximately $50,000 in research in this area yielded a return of over $200,000 per year in documented cost savings (Thresh & Frerichs, 1966). It seems safe to assume that the same kinds of results could be achieved in other organizations.

REFERENCES

LAWSHE, C. H., & TIFFIN, J. The accuracy of precision instrument measurement in industrial inspection. *Journal of Applied Psychology*, 1945, 29, 413–419.

THRESH, J. L., & FRERICHS, J. S. Results through management application of human factors. Paper presented at a meeting of the American Society for Quality Control, New York, June 1966.

(Early publication received January 26, 1967)

Journal of Applied Psychology
1967, Vol. 51, No. 4, 316–319

AFFECT LEVEL, CAPILLARY PULSE PRESSURE, AND RESPONSE LATENCY

BRUCE O. BERGUM AND DONALD J. LEHR

Fundamental Research Laboratory, Xerox, Rochester, New York

An experiment was conducted in which the effects of the interest value of stimuli on pulse rate, capillary pulse pressure, and overt evaluation response latencies were examined. A group of 12 Ss evaluated 20 4-letter words in terms of an interesting-dull dimension under both visual and auditory presentation conditions while the 3 measures of interest were simultaneously and continuously recorded. The results indicated no relationship between pulse rate and interest level, but significant functional relationships between both capillary pulse pressure and overt response latencies and levels of interest. These results substantiated the findings for the latter 2 measures demonstrated in an earlier study which employed pictorial stimuli and a pleasant-unpleasant overt response dimension. The relationships were demonstrated to be unaffected by presentation conditions, and it was concluded that capillary pulse pressure and overt response latencies may discriminate among general affect levels over a broad range of conditions.

Attempts to relate affective states of the individual to objective measures of these states have yielded mixed and sometimes contradictory results. In a comprehensive summary of these results, Cook and Selltiz (1964) have concluded that where physiological measures are concerned, "Most measures of physiological reaction give direct indications only of the extent of arousal; they do not reveal whether the corresponding emotion is pleasurable or unpleasurable [p. 53]." The results of a recent study (Bergum & Lehr, 1966), however, indicate that the affective value of pictorial stimuli, as judged by subjects (*S*s) along a pleasant-unpleasant continuum, can be reliably discriminated in terms both of capillary pulse pressure and overt response latencies as well as in terms of pulse rate and pupil size. In this study, colored photographic stimuli were evaluated on a 5-point scale by means of a keyboard response device, from which the latencies of these evaluative responses were derived, while GSR, capillary pulse pressure, pulse rate, and pupillary dilation were simultaneously and continuously recorded. The results indicated that capillary pulse pressure and pulse rate tended to increase significantly with pleasant stimulation and to decrease with unpleasant stimulation, while response latencies tended to decrease significantly with pleasant stimulation and increase with unpleasant stimulation. Pupil size was demonstrated to be significantly related to affect value, with pleasant stimuli yielding pupillary dilation and unpleasant stimuli yielding pupillary contraction.

The latter results have been demonstrated by other researchers (Hess & Polt, 1960), but the results for capillary pulse pressure, pulse rate, and response latencies appear to be unique in this area, and the purpose of the present study was to attempt to duplicate these results under altered presentation conditions, employing both a different form of stimulation and a slightly different overt response dimension. Specifically, the stimuli for this study were verbal, rather than pictorial, and consisted of four-letter English words ranging from relatively positively affective terms such as HOME and DEAR through such relatively negatively affective terms as SLUT and DEAD. To determine the extent to which the response measures might be affected by stimulus mode, these terms were displayed under both visual and auditory presentation conditions, and in addition, because there was some indication from the subsequent comments of *S*s in the first study that they were responding as much in terms of interest as they were in terms of pleasantness, an interesting-dull rather than the pleasant-unpleasant overt response dimension was employed. These specific references to interest were primarily the basis for selecting this

rather than such alternative dimensions as favorable-unfavorable, for example.

METHOD

Subjects. Twelve females, ranging in age from 20 to 40 yr., served as *Ss* in this study. The sample included both college and high school graduates, married and unmarried women.

Apparatus. The apparatus employed in this study included a rear-projection chamber, a five-key response keyboard, a six-channel event recorder, a Sanborn single-channel continuous paper tape recorder, a pair of stereo headphones, and a Robertson stereo sound tape recorder.

The rear-projection chamber was a closed, soundproofed black box, 18 in. high, 16 in. wide, and 30 in. long, with an opening and chin rest at one end and a 12-in.-square frosted-glass projection screen at the other. This chamber was mounted on a table and the overt response keyboard was located directly to the right side of the chamber on the table. The keyboard consisted of five microswitch detent keys arranged in an arc approximately 4 in. across to conform to the general configuration of the finger pads of the right hand.

Materials and conditions. A total of 41 two-in.-square black and white slides were employed. One of these slides consisted of a white X centered on a medium gray background. Twenty of the slides consisted of a control word, ROCK, presented in white letters on a medium-gray background, and the remaining 20 slides consisted of four-letter stimulus words presented in white letters on a medium-gray background. The 20 stimulus words ranged from terms such as HOME and DEAR, to such terms as DEAD and ACHE. These 20 words were separated into two lists of 10 words each and each slide was paired with a control slide. The pairs of slides were arranged in six different random orders for each list.

In addition to the slide materials, all stimulus and control materials were recorded on stereo sound tape in the same order and time relationships as those employed for the visual materials. The instructions for all conditions were also prerecorded on audio tape.

Two presentation mode conditions, visual and auditory, were employed, with all *Ss* serving under both conditions. The *Ss* were required to fixate on the X slide throughout the auditory presentation condition. The presentation mode conditions, lists, and within-list sequences were counterbalanced across all *Ss*, and, except for this counterbalancing, all *Ss* received identical treatment.

Each control and experimental stimulus was presented for a period of 10 sec., thus requiring a period of 20 sec. for presentation of both members of the pair. At the end of the first presentation mode condition, either visual or auditory, *S* was allowed to relax a few minutes while the changeover was being made for the next condition. The *S* was then instructed by a recording regarding the procedures for the second part of the experiment.

The general procedure consisted of first presenting the control stimulus of a pair, followed by presentation of the experimental stimulus, with *S* performing an overt evaluation response only during the presentation interval for the experimental stimulus. This response was made in terms of a 5-point scale lying along an interesting-dull dimension ranging from *very interesting* through *very dull*. Pulse rate and capillary pulse pressure were continuously recorded throughout the experimental session as were overt responses and response latencies.

RESULTS

The overt evaluation responses served as the criteria for evaluating the three measures of primary interest. In the case of each measure, the data associated with one *interesting*, one *neutral*, and one *dull* randomly selected overt response, in both the visual and auditory conditions, served as the raw scores for the analyses. The results of the analyses for each of the three measures are presented separately.

Pulse rate. This measure consisted of the differences in pulse rate observed between the control and experimental slide presentations of the stimuli, with a constant added to avoid the use of negative numbers. The data thus consisted of relative rather than absolute values of pulse-rate differences.

The mean pulse rates for the three levels of interest category were 2.96 for the interesting category, 3.08 for the neutral category, and 3.00 for the dull category, suggesting no systematic relationship between this measure and the interesting-dull dimension. A repeated-measures analysis of variance was performed on these data in which the effects of presentation conditions, interest level, and *Ss* were compared.

TABLE 1

ANALYSIS OF VARIANCE FOR PULSE RATE

Source	*SS*	*df*	*MS*	*F*
Condition (C)	1	1	1.00	—
Affect (A)	0	2	—	—
Subjects (*Ss*)	11	11	1.00	—
A × C	1	2	.50	1.22
A × *Ss*	19	22	.86	2.10*
C × *Ss*	16	11	1.45	3.54**
Second order	9	22	.41	

* $p < .05$.
** $p < .01$.

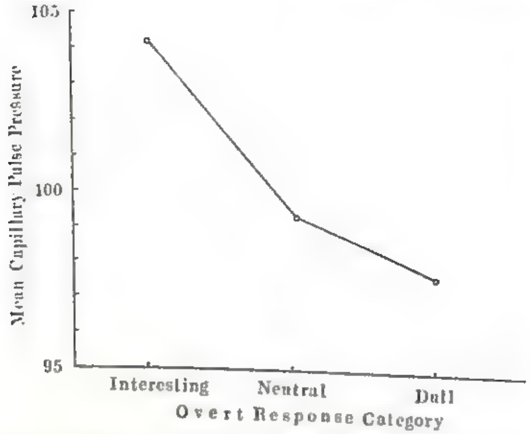


FIG. 1. Mean capillary pulse pressure in millimeters of deflection as a function of overt response category.

The results of this analysis are presented in Table 1. None of the main effects was significant in this analysis. The interaction between *S*s and interest levels was significant ($F = 2.10$, 22 and 22 *df*, $p < .05$) as was that between *S*s and conditions ($F = 3.54$, 11 and 22 *df*, $p < .01$). In general, these results are dissimilar to those demonstrated in the earlier study in which pulse rate was shown to be systematically related to categories of pleasantness.

Capillary pulse pressure. For this measure, the raw data consisted of the mean capillary pulse pressure observed for each stimulus presentation expressed as a percentage of the mean capillary pulse pressure observed for each accompanying control slide presentation.

The mean capillary pulse pressures for the three categories of interest are shown in Figure 1. For this measure, the stimuli evaluated as interesting yielded the highest mean capillary pulse pressures, with the neutral and dull stimuli yielding progressively lower mean pressures. Table 2 gives the results of the repeated-measures analysis of variance for these data similar to that performed on the pulse-rate data. In this analysis, interest category yielded the only significant result ($F = 3.48$, 2 and 22 *df*, $p < .01$). This highly significant effect is similar to that found in the earlier study for this measure. The general magnitude of the effect is indicated by the eta of .314 for this F. Direct-differences *t*'s between the three conditions indicated significant differences between interesting and neutral conditions, $t(11) = 2.435$, $p < .05$, and between the interesting and dull conditions, $t(11) = 3.604$, $p < .01$. The difference between neutral and dull conditions was not significant, $t(11) = 0.310$, $p > .05$, however.

Response latency. This measure was recorded in terms of millimeters of event recorder tape traversed from the point of initiation of the stimulus to the point at which an overt response was performed. The data were then subjected to a log transformation and these transformed data served as the raw scores for the subsequent analysis.

Figure 2 illustrates the mean log latencies plotted as a function of overt response category. In general, response latencies appear to be shortest for the stimuli evaluated as interesting, and somewhat longer but approxi-

TABLE 2

ANALYSIS OF VARIANCE FOR PULSE PRESSURE

Source	*SS*	*df*	*MS*	*F*
Condition (C)	1	1	1.00	—
Affect (A)	424	2	212.00	3.48*
Subjects (*Ss*)	408	11	37.09	—
A × C	84	2	42.00	—
A × *Ss*	850	22	38.64	—
C × *Ss*	605	11	55.00	—
Second order	1,928	22	87.64	
Pooled error term for main treatment effects	3,467	57	60.82	

* $p < .05$.

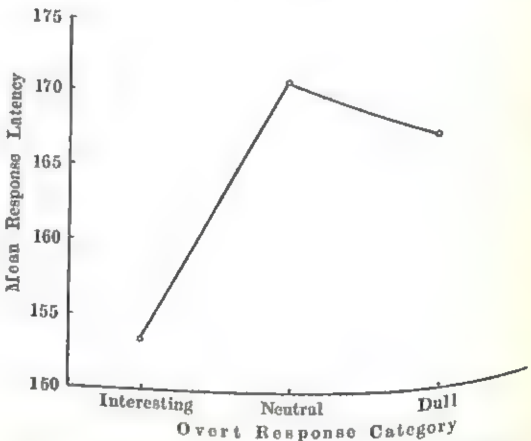


FIG. 2. Mean overt response latencies in arbitrary units of response tape as a function of overt response category.

mately the same for the neutral and dull stimuli. A repeated-measures analysis of variance of these data is given in Table 3. The results indicate that the effects of interest category were highly significant ($F = 8.42$, 2 and 22 *df*, $p < .01$), as was the case for presentation condition ($F = 20.84$, 1 and 22 *df*, $p < .01$). In the case of interest category, the extent of the relationship is indicated by the obtained eta of .354. In addition, the differences between *S*s were also highly significant ($F = 3.59$, 11 and 22 *df*, $p < .01$). Direct-differences *t*'s between the three interest levels indicated that the difference between interesting and neutral conditions was significant, $t(11) = 4.695$, $p < .01$, as was that between the interesting and dull conditions, $t(11) = 4.190$, $p < .01$. The difference between the neutral and dull condition was not significant, $t(11) = 0.852$, $p > .05$. These results are very similar to those demonstrated in the earlier study where a pleasant-unpleasant overt response dimension was employed.

TABLE 3

ANALYSIS OF VARIANCE FOR RESPONSE LATENCY

Source	*SS*	*df*	*MS*	*F*
Condition (C)	5,992	1	5,992.00	21.09*
Affect (A)	4,785	2	2,392.50	8.42*
Subjects (*Ss*)	11,224	11	1,020.36	3.59*
A × C	983	2	491.50	1.79
A × *Ss*	4,310	22	195.91	—
C × *Ss*	4,863	11	442.09	1.61
Second order	6,039	22	274.50	
Pooled error term for main treatment effects	16,195	57	284.12	

* $p < .01$.

DISCUSSION

The present results indicate that two of the three objective measures employed in this study, capillary pulse pressure and overt response latencies, discriminate reliably between different levels of expressed interest, while the third measure, pulse rate, does not. With the exception of the latter result, these findings duplicate almost exactly those demonstrated in an earlier study (Bergum & Lehr, 1966) in which a pleasant-unpleasant overt response dimension was employed and suggests a common underlying mechanism for both the pleasant-unpleasant and interesting-dull dimensions.

The results also suggest that the precise form of stimulation does not alter the fundamental relationship between capillary pulse pressure and overt response latencies and what might be called positive and negative, approach and avoid, or affective categories of stimulation, since in one study complex colored pictorial stimuli were employed, while in the second study simple verbal stimuli were employed under both visual and auditory presentation conditions. Thus it seems reasonable to eliminate stimulus complexity as an explanatory mechanism for the observed results in these studies.

The significantly longer latencies observed under the auditory presentation conditions (172.6 as opposed to 154.4 for the visual) were to be anticipated simply as a function of the relatively longer transmission times required for the auditory display of information as compared to its display in visual form.

Because this experiment in a sense constitutes a replication of the earlier study, the results contribute considerably, both in a statistical and a heuristic sense, to the confidence that can be placed in the observed relationships and imply that these measures might profitably be employed in a broad range of situations involving a variety of affective stimulation of the kinds described by Cook and Selltiz (1964) in their summary of research on attitude measurement.

REFERENCES

BERGUM, B., & LEHR, D. J. *Prediction of stimulus approach: Core measures experiment I.* Rochester, N. Y.: Xerox Corporation, 1966. (Research Report R66-8)

COOK, S. W., & SELLTIZ, C. A multiple-indicator approach to attitude measurement. *Psychological Bulletin,* 1964, **62**, 36–55.

HESS, E. H., & POLT, J. M. Pupil size as related to interest value of visual stimuli. *Science,* 1960, **132**, 349–350.

(Received July 26, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 4, 320–323

RELATIONSHIP BETWEEN FAMILY BACKGROUNDS AND WORK VALUES

FRANK T. PAINE, DONALD R. DEUTSCH, AND RODNEY A. SMITH

Department of Business Administration, University of Maryland

Previous evidence has indicated that work values play an important role in vocational choice and job satisfaction and begin to stabilize in adolescence. This questionnaire study further investigated the relationship between family-background characteristics and expected work values for college undergraduates. Different, distinct, family-background patterns, e.g., including family income, number of times moved, emphasis on discipline, emphasis on material things, closeness of family, etc., existed for those youths who stress humanitarian service (helping others); those who stress security and pleasant associations and working conditions; those who emphasize prestige, responsibility, and independence; those who stress creativity; and those who emphasize monetary benefits.

The family has long been recognized as an important factor in vocational development. In a study of work adjustment, it was concluded that people take their early experiences and attitudes with them on their jobs and react to their work accordingly (Friend & Haggard, 1948). Centers (1948) found significant differences in work values among persons from various social strata: the higher groups emphasized the opportunity for self-expression and leadership; the lower groups stressed security and the desire for independence. Porter (1954) found the prestige level of the father's occupation was related to vocational planning and preference of high school boys. MacArthur (1955) found significant differences in attitudes toward work between college students from upper- and from middle-class homes. Work-value differences that have been shown between public- and private-school graduates may also be in part attributed to differences in orientation in the family (Wilson, 1959). Furthermore, these values toward work tend to stabilize in adolescence (Ginzberg, Ginzburg, Axelrad, & Herma, 1951).

Still needing further exploration, however, is the question of which family-background characteristics are associated with which work values (Kluckhohn & Strodtbeck, 1961; Norton & Kuhlen, 1950; Super, Crites, Hummel, Moser, Overstreet, & Warnath, 1957). To this end such a comparison was made to add to the understanding of vocational choice and job satisfaction.

METHOD

The data for this study were obtained by means of a questionnaire containing three parts: the importance of work values, the perception of family-background characteristics, and general family information.

Work values were defined as the perception an individual has of the factors associated with a job which satisfy his needs and help him reach his goals. Nine work factors were used: job security, physical working conditions, social needs (close and pleasant associations with fellow workers and opportunity to help others), job prestige, high degree of responsibility, opportunity for independent action, opportunity to invent or design new things or ideas, and monetary compensation and employee benefits. The respondents were asked, "When you begin working, how important do you think the following characteristics will be to you?" For each of the work values they were instructed to answer by circling a number on a rating scale extending from 1 to 7, where "low numbers represent low or minimum amounts and high numbers represent high or maximum amounts."

Similarly, the respondents were asked to rate on a 7-point scale the amount of certain family-background factors perceived to be present in their upbringing, that is,

> how close were the members of your family, how much discipline were you subject to during your youth, how many close friends did you have as a child, how much did your family encourage you to be independent, and during your youth to what degree were the following factors emphasized in your home: religion, cultural interest, materialistic things, community activities (i.e., Boy Scouts, youth groups, etc.), social activities (dating, parties, etc.).

TABLE 1

SIGNIFICANT CORRELATIONS BETWEEN WORK VALUES AND FAMILY BACKGROUND

Family background	Work values								
	Security	Physical working conditions	Pleasant associations with fellow workers	Help others	Prestige	Responsibility	Independence	Creativity	Monetary benefits
Closeness	.32***	.25**	.16*	—	—	.15*	—	−.15*	
No. close friends	.31***	.26**	.32***	—	—	.17*	—	—	.19**
Discipline	.30***	.28***	.31***	—	—		—	−.15*	.30***
Social activities	.23**	.21**	.19**	.17*	.39***	.28***	.23**	—	—
Community activities	.20**	.22**	.21**	—	.22**	.24**		−.17*	.19**
Religion	.20**	—	.17*	—	—	—		—	.21**
Independence	—	—	—	—	.17*	.33***	.31***	—	
Culture	—	—	—	.24**	.18*	.27***	.20**	—	
Material things	—	.30***	—		.39***	—	—	—	.15*
Marital status of parents	—	—	—	—	—	—	—	—	—
No. children in family	—	—	—	—	—	—	—	—	—
No. times moved	—	—	—	—	−.25**	−.23**	−.14*	—	—
Family income	−.22**	—	—	—	—	—	—	—	.21**

Note.—$N = 145$.
* $p < .05$, $r \geq .137$.
** $p < .01$, $r \geq .192$.
*** $p < .001$, $r \geq .269$.

General information was gathered on family income, number of children in family, the number of times the family moved, and marital status of parents.

The questionnaire was administered to 155 male University of Maryland undergraduates. They were mostly freshmen and sophomores and were from a cross section of major departments and schools. A Pearsonian correlation computer program was used to statistically compare the relationships from 145 usable questionnaires.

RESULTS

Table 1 indicates the significant correlations between work values and family-background factors. Each work value is associated with at least two or up to seven perceived family-background factors. The correlations, however, though statistically significant, are usually low or moderate.

The table also indicates the pattern of certain background factors. For example, family social activities has moderate positive correlations with prestige, responsibility, and independence. The number of times moved, on the other hand, has negative correlations with the same values. Family income shows one positive correlation (.21 with monetary benefits) and one negative correlation (−.22 with the security need). Moreover, the grouping of correlations indicates various family-background patterns relating to various job orientations.

THE ENVIRONMENTALIST

A pattern of six moderate positive (.20–.32) correlations and the one negative correlation existed for the security need. As might be expected, those respondents who indicated a relatively low family income tended to place a high emphasis on the job-security value. This supports the Centers' study (although his sample was made up of people in their working years).

Furthermore, the most effective combination for predicting security emphasis would also include those who indicate a close family relationship with discipline, social activities, community activities, and religion emphasized, and a number of close friends seen as existent. This type of family perhaps "makes up" for lack of income by imbuing a social-religious security which affects the developmental process of work values.

The pattern of responses for the physical-working-conditions factor was similar to that for security with two exceptions; ones that might be expected: An emphasis on material things in the home associated with the importance of pleasant working conditions; whereas no significant correlation existed between religious emphasis and pleasant working conditions.

The social need for close association with others on the job also follows a similar pat-

tern of correlation with background factors with the exception of the income item (where no significant correlation exists). The work group thus may provide important security and camaraderie for a person with a family background as indicated.

Moreover, speculation on this similarity in the three patterns of family-background responses for those who value security, pleasant working conditions, and associations might indicate an "environmentalist" type of vocational development, that is, one who emphasizes a pleasant, secure environment.

The Humanitarian

Another social need (helping others) shows a different pattern and it is more difficult to predict with only two significant correlations with family background. The highest correlation (.24) is with perceived emphasis on culture in the home. The other correlation, a relatively low one (.17), is with social activities emphasis. These results may seem hard to explain; one might expect, for example, for religion (or other factors not measured) to enter in. Perhaps, however, the cultural stimulation was developed with congenial interpersonal relations, both inside and outside the home, and an "other awareness" developed. Thus, one might surmise the "humanitarian" vocational outlook for youths of this background.

The Climber

The work values of prestige, responsibility, and independence indicate similar relationships with several family-background factors, that is, emphasis on social activities, independence, and culture, and the number of times the family moved. In addition, prestige alone (of the three) was correlated (.39) with the emphasis on material things. Also, prestige and responsibility associated with community activities.

From this sample, it can be speculated that the amount of prestige desired by one selecting a vocation is a function of the extent he was encouraged to participate in social activities and the emphasis on material things in his home life. Similarly, the evidence suggests that the youth who has been brought up in a family where culture, independence, and social activities have been emphasized would tend to explore vocations where responsibility and independent action are perquisites. On the other hand, frequent family relocations during youth are associated with less emphasis on job prestige, responsibility, and independence.

Thus, the somewhat similar family-background patterns for those who stress these three work values might indicate the growth and exploratory stage for the vocational "climber."

The Creator

No support existed for the possible effect of income (social strata) on self-expression (and independence) as the Centers' study indicated. In fact, the creativeness job characteristic was associated only with low negative correlations with the amount of community activities, discipline, and closeness of family. If one assumes that individuals who value creativity tend to be creative, this evidence provides modest support for other studies (see Chambers, 1964; MacKinnon, 1965; Scott, 1965) which have indicated less discipline, less closeness to parents, and fewer rewards from family for community interaction for creative individuals. On the other hand, there was no support for the previous finding of a positive association of moving about in childhood and creativity.

The Mercenary

Finally, the monetary job factor was associated with several family-background factors. Most prominent was the amount of discipline with a moderate positive correlation (.30); of lesser relationship was the amount of income and emphasis on community activities, the number of close friends, material things, and, perhaps surprisingly, religion. The materialistic, well-to-do, disciplined family which seems to emphasize religion and aspire for close friends may affect its offspring toward a greater emphasis on monetary compensation and benefits.

Results of this correlational study though do not warrant unlimited generalization. Work

orientations no doubt are influenced by a number of interacting factors in the family subculture; more evidence is needed to check the several patterns of development that are suggested by the data.

REFERENCES

CENTERS, R. Motivational aspects of occupational stratification. *Journal of Social Psychology*, 1948, **28**, 187–217.

CHAMBERS, J. A. Relating personality and biographical factors to scientific creativity. *Psychological Monographs*, 1964, **78**(7, Whole No. 584).

FRIEND, J. G., & HAGGARD, E. A. Work adjustment in relation to family background. *Applied Psychological Monographs*, 1948, No. 16.

GINZBERG, E., GINZBURG, S. W., AXELRAD, S., & HERMA, J. L. *Occupational choice: An approach to a general theory*. New York: Columbia University Press, 1951.

KLUCKHOHN, F. R., & STRODTBECK, F. L. *Variations in value orientations*. Evanston, Ill.: Row, Peterson, 1961.

MACKINNON, D. W. Personality and the realization of creative potential. *American Psychologist*, 1965, **20**, 273–281.

MCARTHUR, C. Personality differences between middle and upper classes. *Journal of Abnormal and Social Psychology*, 1955, **50**, 247–254.

NORTON, J. L., & KUHLEN, R. G. The development of vocational preferences as revealed in vocational histories obtained by interview. *American Psychologist*, 1950, **5**, 296–302.

PORTER, J. R. Predicting vocational plans of high school senior boys. *Personnel and Guidance Journal*, 1954, **33**, 215–218.

SCOTT, W. E. The creative individual. *Academy of Management Journal*, 1965, **3**, 211–219.

SUPER, E. E., CRITES, J. O., HUMMEL, R. C., MOSER, H. P., OVERSTREET, P. L., & WARNATH, C. R. *Vocational development: A framework for research*. New York: Columbia University Press, 1957.

WILSON, W. C. Value differences between public and private school graduates. *Journal of Educational Psychology*, 1959, **50**, 213–218.

(Received August 1, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 4, 324–329

MOTIVATIONAL EFFECTS OF KNOWLEDGE OF RESULTS: KNOWLEDGE OR GOAL SETTING? [1]

EDWIN A. LOCKE

American Institutes for Research, Washington Office

Research on knowledge of results (KR) has generally not controlled for motivational effects resulting from differential goal setting. The present experiment was carried out to separate the effects of KR and goal setting using a 2 × 2 fixed-model design; the variables were KR versus no KR; and specific hard goals versus "do-best" goals. The goals (manipulated by instructions) were representative of the goals typically assigned (explicitly or implicitly) to KR and No KR Ss in previous studies, respectively. No difference was found between KR and No KR groups, but a significant goal effect was found in favor of Ss given specific hard goals. The results indicated that effects previously attributed to differential KR were actually due to different levels of motivation produced by the different goals.

The positive influence of knowledge of results (KR) on learning and performance is one of the best established findings in the research literature (Ammons, 1956; Bilodeau & Bilodeau, 1961, etc.). Research on KR has not been overly concerned with theoretical issues, but the fact that occasional studies fail to find significant KR effects on performance (e.g., Chapanis, 1964) emphasizes the need to integrate and systematize findings in this area.

Several investigators (e.g., Vroom, 1964) have distinguished between the *information* or cueing function of KR and the *motivational* function. The former refers to information given to *S*s regarding the nature and locus of errors and the nature of the correct response (epistemic KR). Most studies of KR or "feedback" have been of this type (e.g., Bilodeau & Bilodeau, 1961). Given a constant motivational state, the more information given to *S* about the task or about how to correct errors, the better his performance level or learning.

The manner in which knowledge functions to *motivate S* is much more complicated. One possible means of isolating the motivational effect of KR is to give *S* knowledge that cannot be used for correcting errors or for changing the timing or locus of the response. Examples of KR that do not cue *S* as to a better *method* of performing would be: total time-on-target on a pursuit rotor task; total score after several trials on a computation task; total score after several throws on a dart-throwing task. Knowledge of score on simple motor tasks such as reaction time and weight lifting also falls into this category, since there are no right or wrong responses on such tasks. Any effects on performance of this type of KR may be attributed to motivational factors since presumably only *S*'s level of effort is influenced.[2]

In the "classic" studies dealing with motivational KR such as those of Book and Norvell (1922), Crawley (1926), Mackworth (1950), and Ross (1927), *S*s given knowledge of score were usually told to *try to improve their performance* or were given explicit goals to reach, whereas the No KR *S*s were told not to think about trying to improve their scores or were told simply to "do their best." Although the KR groups performed better than the No KR groups in these experiments, the effects of goal setting were confounded with the effects of knowledge of score, making it

[1] This research was supported by contract Nonr 4792(00) between the Office of Naval Research and the American Institutes for Research. The author would like to thank Judith Bryan of AIR for analyzing the data for this study and Charles D. Spielberger of Florida State University for his many helpful comments on this paper.

[2] While it is true that scores of this type could be taken as a cue or signal to *change* one's method of performing the task or to change one's strategy, such scores do not yield information about *what* one did wrong or *where* the mistake(s) occurred. Nor does such knowledge tell one what *kinds* of corrections to make.

impossible to determine which was the critical variable or the interdependence, if any, between them. Similar criticisms can be made of recent studies, for example, Church and Camp (1965); McCormack, Binding, and Chylinski (1962); McCormack, Binding, and McElheran (1963); and Payne and Hauty (1955). In these studies goals were not manipulated explicitly, but KR was given in relation to standards (such as S's best previous score) which clearly must have suggested goals to Ss.

In five other studies (Arps, 1920; Gibbs & Brown, 1955; Johanson, 1922; Mace, 1935; Manzer, 1935) positive results were obtained for KR without explicit manipulation of Ss' goals; however, no attempt was made to determine if different goals were set by KR and No KR groups (although Mace argued that the effect of KR was to suggest performance standards to the KR Ss). In contrast, Chapanis (1964), attempting to replicate the finding of Gibbs and Brown (1955), did two things to eliminate "demand characteristics" (Orne, 1962) that may have affected previous results: (*a*) he "hired" Ss as employees doing a job for pay (punching digits onto a tape) rather than as experimental Ss; and (*b*) he ran Ss individually rather than in groups (as was the case in most previous experiments in this area). In the absence of the implicit goal-setting demands (e.g., "improvement") inherent in the typical experimental situation and the possible effects of implicit competition, Chapanis found no effect of KR on performance. Further, Locke and Bryan (1966b) found no overall effect of KR on performance on a complex computation task. But when Ss were regrouped according to their a posteriori performance goal descriptions, a significant relationship of goals to performance was found, suggesting that differences in Ss' goals had more effect on performance than differences in KR.

The above findings suggest that the effects of KR should be separated from the effects of goal setting in order to determine whether KR influences performance level automatically or whether its effects are mediated by goal-setting activity. The purpose of the present experiment was to test the hypothesis that effects previously attributed to differential KR were actually due to different performance goals associated with the different KR conditions. The major premise underlying this hypothesis is that level of effort on a task is determined largely by S's conscious performance goals. This emphasis on a cognitive approach to motivation is supported by recent theoretical developments (e.g., Dulany, 1962; Miller, Galanter, & Pribram, 1960; Ryan, 1958[3]; Spielberger, 1965) and by the findings of a number of recent studies (e.g., DeNike, 1965; Dulany, 1962; Locke, 1966a, 1966b; Locke & Bryan, 1966a, 1966b; Spielberger, Berger, & Howard, 1963; Spielberger, Bernstein, & Ratliff, 1966; Spielberger, Levin, & Shepard, 1962).

METHOD

Task

The task was simple addition. Each problem consisted of three two-digit numbers and was presented on a separate 3 × 5 index card. The cards were placed consecutively in boxes holding 720 cards each. The Ss wrote their answers on sheets containing space for 90 answers. As each answer sheet was completed, S was instructed to insert it in a slot underneath a one-way mirror through which he was observed by *E* during the experiment. The S worked 1 hr. at the task, which was divided into five trials separated by short rest periods. The trials were alternately 10 and 15 min. in length (i.e., 10′; 15′; 10′; 15′; and 10′). All Ss were told the length of each trial in advance.

Subjects

The Ss were 24 male and 12 female paid college (undergraduate) volunteers selected from a larger pool of 62 volunteers. All Ss in the original pool were given a pretest consisting of three 1-min. trials on the addition task and asked to make three attitude ratings indicating their liking for and interest in "tasks like this." From this pool, four matched groups of nine Ss each were selected as follows: four Ss were chosen who were approximately equal in ability and equal on the three attitude ratings; each of these Ss was then assigned at random to one of the four experimental conditions. This procedure was repeated until there were nine Ss in a cell. Thus the cells had almost identical means and distributions on the four matching variables. No attempt was made to match for sex; but as it turned

[3] See also unpublished mimeos, 1964. Chapter I: Explaining behavior; Chapter II: Explanatory concepts; Chapter V: Experiments on intention, task, and set; Chapter VI: Intentional learning; Chapter VII: Unintentional learning. Cornell University, Department of Psychology.

out, the proportion of females to males was approximately equal for the main effects.

Design and Procedure

The experiment was introduced as a study of the development of attitudes toward the task. The structure of the task was explained. The *S*s were told not to spend a lot of time checking answers since the total number correct was more important than the percentage correct. While *S* worked, *E* stayed in the adjacent observation room. The *S* was made aware of the one-way mirror and was told *E* would remain there during each trial so as not to disturb *S*. Communication was possible through an intercom system.

The design was a 2×2 fixed model. The fixed variables were knowledge of score (KR) and type of goal.

KR condition. The KR *S*s were told the number of problems they had gotten correct at the end of each trial.

No KR condition. The No KR *S*s were not given their scores. Since *S*s had to hand in each answer sheet after completion and since they did not know how many of their answers were wrong, these *S*s could not easily keep track of their scores.

Do-Best goal. Do-Best *S*s were told to do their best on each trial. (This was the goal typically given to No KR *S*s in the experiments discussed earlier.) Do-Best *S*s in the KR condition could not easily use their scores to set specific personal goals due to the alternating trial lengths. In addition, Do Best-KR *S*s were not allowed to perform any computations between trials that might enable them to determine their rate per minute (a precondition for setting a goal such as "improvement").

Hard goal. Goals were set for the Hard-Goal *S*s on the basis of the scores attained by the matched Do-Best *S*s. On the first trial a given *S*'s goal was set about 10% higher than the score achieved on the same trial by a matched Do-Best *S*. Then *E* adjusted *S*'s goal before each succeeding trial depending upon how well *S* had done on the previous trial. If *S* did not get near the goal, the next goal was lowered slightly; if *S* reached or exceeded the goal, the succeeding goal was raised.

The goals were marked by means of a colored 3×5 index card placed vertically at the appropriate point in the box of problem cards. This card represented the point they had to reach by the end of the trial in order to reach their goal (if they got all problems correct). The *S*s were told to try and surpass this point since they were bound to get some wrong. They were told at the end of each trial whether or not they had beaten the goal but were not given their actual scores unless they were in the KR condition.

On the average the goals of the Hard-Goal *S*s were set from 0% to 32% (mean and median = 11%) above the scores attained by the matched Do-Best *S*s. Hard-Goal *S*s were actually able to reach or beat their goals on 16% of the trials.

The *S*s had approximately 2 min. rest between trials during which *E* corrected their answer sheets (if they were in the KR condition) and determined their new goals (if in the Hard-Goal condition). To set the new goal, *E* came in the experimental room and placed the colored card at the appropriate place in the card box. No *S*s were told how many problems were needed to reach the goal, and they could get only a rough idea of this by seeing where the colored card was placed in the card box.

At the end of the experiment *S*s were asked to look at a set of cards, each of which described a possible goal they could have had (e.g., "I tried to do my best"; "I tried for the assigned goal") and were asked to pick out the card that best represented their performance goal during the experiment.

RESULTS

Three different performance measures were used: (*a*) deterioration scores, defined as the difference between the mean number of problems correct per minute on the three pretest trials and the mean number of problems correct per minute during the five experimental trials; (*b*) linear slope scores calculated from the number of problems correct per minute on the five experimental trials only; and (*c*) percentage of errors on the five experimental trials.

The performance curves for each experimental subgroup, in terms of total problems correct per minute, are shown graphically in Figure 1.[4] It is evident that both Hard-Goal groups did progressively better than the Do-Best groups, the difference between them reaching a peak in the last (10′) trial period. On the other, the KR and No KR curves did not diverge except for the Hard-Goal group in the last 10-min. period.

The results of analyses of variance (using a standard 2×2 fixed-model design) for the three performance measures are shown in Table 1. The goal effect was significant using both deterioration scores, $F\ (1/32) = 4.83$,

[4] It is evident that, in general, *S*s showed deterioration (rather than improvement) in their rate of performance from the pretest to the experimental trials. This can be attributed to two factors. First, little or no learning occurs over a short period of time with mature *S*s on a task like addition. Second, the slightest lapse in concentration will slow one's addition rate. Since over a longer time span some attention lapses are inevitable, this produces a slower rate on the longer trials. The experimental trials in this case were 10–15 times as long as the (three 1-min.) practice trials.

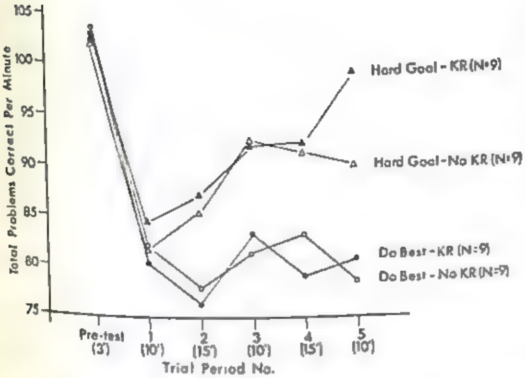


Fig. 1. Total problems correct by trial period for experimental subgroups.

$p < .05$, and linear slope scores, $F\ (1/32) = 12.87$, $p < .01$, with the Hard-Goal group being superior to the Do-Best group in both cases. A matched-groups test comparing the Hard-Goal and Do-Best groups (performed by matching each Hard-Goal *S* with a Do-Best *S* of equal initial ability) yielded a *t* ratio for mean deterioration of 2.80 (17, $p < .02$), and a similar test using slope scores yielded a *t* of 3.18 (17, $p < .01$). There was no overall effect of KR on any of the performance measures, nor was there evidence of any interaction between KR and type of goal.

There were no significant main effects nor any interaction for the percent error criterion, indicating that the higher performance level attained by the Hard-Goal group was not achieved at the expense of a relatively greater number of errors in relation to problems attempted. Both the Hard-Goal and Do-Best groups averaged about 5.6% errors during the experimental trials.

The results of the postexperimental questioning indicated that all Hard-Goal *S*s were either trying explicitly for the goals or had them in mind as they worked along. Eleven of the Do-Best *S*s said they were trying to "do their best"; four said they were trying to go at a "reasonably fast pace"; two tried to go at a "normal" or "slow" pace; and one *S* reported that he tried to complete one box of (720) problem cards by the end of the hour. The latter *S* was the only *S* in the Do-Best condition who worked at a faster pace (in terms of problems correct per minute) during the experimental than during the three pretest trials.

Discussion

The results of the present experiment support the hypothesis that motivational effects previously attributed to differential KR were actually a function of differential performance goals associated with the KR conditions. When differential goal setting by KR and No KR *S*s was controlled, no overall effect of KR on performance was found. However, when the effects of the goals typically set by KR and No KR *S*s in previous experiments (e.g., specific hard goals versus do best) were compared, a significant goal effect was found. These results were consistent with previous findings by Locke and Bryan (1966b) where goals were measured by postexperimental interviews rather than manipulated by instructions.

Although the findings of the present study provide support for the initial hypothesis, there are a number of issues still open. It may be observed from Figure 1 that there was an effect of KR condition between Trials 4 and 5. The KR groups increased their performance rate on Trial 5, whereas the No KR groups decreased. The *F* for the difference scores

TABLE 1

Analysis of Variance Results for Three Performance Criteria

Source	*F*
Deterioration	
Goals	4.83*
KR	<1
Goals × KR	<1
Within (*MS*)	(8,695.28)
Linear slope (Trials 1–5)	
Goals	12.87**
KR	1.36
Goals × KR	<1
Within (*MS*)	(797.78)
Percentage of errors (no. wrong/no. attempted)	
Goals	<1
KR	1.32
Goals × KR	<1
Within (*MS*)	(9.29

Note.—$df = 1/32$.
* $p < .05$.
** $p < .01$.

(1/32) was 8.83 ($p < .01$). It would be useful to determine whether this effect was due to implicit goal setting on the part of the KR *S*s, who by this time may have been able to get some idea of their rate, or to some other effect.

Only two goal classes were used in the present study. In one sense the hard goals in the present experiment were harder than the do-best goals, since the hard goals were set above the performance level of the Do-Best *S*s. However, just "how much" harder is not known. A greater variety of performance goals should be utilized in subsequent studies and some attempts made to scale the goals as to difficulty or motivational level.

In more general terms, the results of the present study suggest one mechanism by which incentives of all types might work (e.g., KR, money, instructions, participation, praise, reproof, verbal "reinforcement," etc.); it is possible that such incentives are effective only to the degree that they affect *S*'s goals or intentions. Research in the area of verbal learning (DeNike, 1965; Dulany, 1962; Spielberger et al., 1962) suggests that a verbal reinforcement (given that the response-reinforcement contingency is known) is effective in changing behavior to the precise extent that *S* desires or intends to get the reinforcement. In view of the seemingly inconsistent and often unpredictable effects of incentives on performance, it seems probable that the use of goals and intentions as mediating variables will be necessary if fully adequate explanations of the effects of incentives are to be achieved.

The present results suggest that in order to predict the effect of knowledge of score on performance level, it is not enough to know *that* the individual has such knowledge. It is also necessary to know *what he does with it,* that is, how he evaluates it and what goals he sets in response to it.

REFERENCES

AMMONS, R. B. Effects of knowledge of performance: A survey and tentative theoretical formulation. *Journal of General Psychology,* 1956, **54**, 279–299.

ARPS, G. F. Work with knowledge of results versus work without knowledge of results. *Psychological Monographs,* 1920, **28**(3, Whole No. 125).

BILODEAU, E. A., & BILODEAU, I. McD. Motor-skills learning. *Annual Review of Psychology,* 1961, **12**, 243–280.

BOOK, W. F., & NORVELL, L. The will to learn: An experimental study of incentives in learning. *Pedagogical Seminary,* 1922, **29**, 305–362.

CHAPANIS, A. Knowledge of performance as an incentive in repetitive, monotonous tasks. *Journal of Applied Psychology,* 1964, **48**, 263–267.

CHURCH, R. M., & CAMP, D. S. Change in reaction time as a function of knowledge of results. *American Journal of Psychology,* 1965, **78**, 102–106.

CRAWLEY, S. L. An experimental investigation of recovery from work. *Archives of Psychology,* 1926, No. 85.

DENIKE, L. D. Recall of reinforcement and conative activity in verbal conditioning. *Psychological Reports,* 1965, **16**, 345–346.

DULANY, D. E., JR. The place of hypotheses and intentions: An analysis of verbal control in verbal conditioning. In C. W. Eriksen (Ed.), *Behavior and awareness.* Durham, N. C.: Duke University Press, 1962. Pp. 102–129.

GIBBS, C. B., & BROWN, I. D. Increased production from the information incentive in a repetitive task. Medical Research Council, Applied Psychology Research Unit (Great Britain), 1955, No. 230.

JOHANSON, A. M. Influence of incentive and punishment on reaction time. *Archives of Psychology,* 1922, **8**, No. 54.

LOCKE, E. A. A closer look at level of aspiration as a training procedure: A reanalysis of Fryer's data. *Journal of Applied Psychology,* 1966, **50**, 417–420. (a)

LOCKE, E. A. The relationship of intentions to level of performance. *Journal of Applied Psychology,* 1966, **50**, 60–66. (b)

LOCKE, E. A., & BRYAN, J. F. Cognitive aspects of psychomotor performance: The effects of performance goals on level of performance. *Journal of Applied Psychology,* 1966, **50**, 286–291. (a)

LOCKE, E. A., & BRYAN, J. F. The effects of goal-setting, rule-learning, and knowledge of score on performance. *American Journal of Psychology,* 1966, **79**, 451–457. (b)

MACE, C. A. Incentives: Some experimental studies. Industrial Health Research Board (Great Britain), 1935, Report No. 72.

MACKWORTH, N. H. Researches on the measurement of human performance. Medical Research Council Special Report Series (Great Britain), 1950, No. 268.

MANZER, C. W. The effect of knowledge of output on muscular work. *Journal of Experimental Psychology,* 1935, **18**, 80–90.

MCCORMACK, P. D., BINDING, F. R. S., & CHYLINSKI, J. Effects on reaction-time of knowledge of results of performance. *Perceptual and Motor Skills,* 1962, **14**, 367–372.

MCCORMACK, P. D., BINDING, F. R. S., & MCELHERAN, W. G. Effects on reaction-time of partial

knowledge of results of performance. *Perceptual and Motor Skills,* 1963, **17**, 279–281.

MILLER, G. A., GALANTER, E., & PRIBRAM, K. H. *Plans and the structure of behavior.* New York: Holt, 1960.

ORNE, M. T. On the social psychology of the psychological experiment with particular reference to demand characteristics. *American Psychologist,* 1962, **17**, 776–783.

PAYNE, R. B., & HAUTY, G. T. Effect of psychological feedback upon work decrement. *Journal of Experimental Psychology,* 1955, **50**, 342–351.

ROSS, C. C. An experiment in motivation. *Journal of Educational Psychology,* 1927, **18**, 337–346.

RYAN, T. A. Drives, tasks, and the initiation of behavior. *American Journal of Psychology,* 1958, **71**, 74–93.

SPIELBERGER, C. D. Theoretical and epistemological issues in verbal conditioning. In S. Rosenberg (Ed.), *Directions in psycholinguistics.* New York: Macmillan, 1965. Pp. 149–200.

SPIELBERGER, C. D., BERGER, A., & HOWARD, K. Conditioning of verbal behavior as a function of awareness, need for social approval, and motivation to receive reinforcement. *Journal of Abnormal and Social Psychology,* 1963, **67**, 241–246.

SPIELBERGER, C. D., BERNSTEIN, I. H., & RATLIFF, R. G. Information and incentive value of the reinforcing stimulus in verbal conditioning. *Journal of Experimental Psychology,* 1966, **71**, 26–31.

SPIELBERGER, C. D., LEVIN, S. M., & SHEPARD, M. The effects of awareness and attitude toward the reinforcement on the operant conditioning of verbal behavior. *Journal of Personality,* 1962, **30**, 106–121.

VROOM, V. H. *Work and motivation.* New York: Wiley, 1964.

(Received August 3, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 4, 330–339

THE HERZBERG THEORY:

A CRITIQUE AND REFORMULATION

CARL A. LINDSAY,[1] EDMOND MARKS,[2] AND LEON GORLOW

Pennsylvania State University

A formal statement of the Herzberg Theory was undertaken which specified functional relations between the variables of interest. A nonadditive model of job satisfaction was advanced to test the relationships. A complete $3 \times 3 \times 2$ factorial design (Motivators × Hygienes × Employee Classification) was employed to test 3 hypotheses concerning the variance accounted for by the treatment effects and the relationship of the independent and dependent variables. Treatment effects were varied by presenting 18 groups of 15 industrial workers each with pairings of different qualitative levels of job factors in a questionnaire format. The 3 hypotheses were supported by the results, and it was concluded that Herzberg's conception of job satisfaction as being comprised of 2 unipolar continua should be reevaluated.

The Herzberg Theory of job satisfaction, first published in 1959 (Herzberg, Mausner, & Snyderman, 1959), and amplified and extended by Herzberg (1962, 1964; Herzberg & Hamlin, 1961, 1963), represents an approach to an understanding of the motivation to work. In the original (1959) study, the investigators concluded that the determinants of job satisfaction, motivators, are different from the determinants of job dissatisfaction, hygienes. Motivators which are job-content variables are called such to emphasize their role in satisfying the worker's need for self-actualization. Job-context variables are labeled hygienes to indicate their role in preventing job dissatisfaction. Herzberg et al. (1959) further concluded that motivators affect job attitudes in a positive direction, and "their absence will much less frequently lead to job dissatisfaction [pp. 81–82]." On the other hand, hygienes "represent the major job dissatisfiers with little potency to affect job attitudes in a positive direction [pp. 81–82]." In a later theoretical paper, Herzberg (1964) took the next logical step and decided that job satisfaction and job dissatisfaction were not the opposite of each other, but . . . "that job satisfaction is made up of two unipolar traits [p. 3]."

By now the theory has received rather widespread attention in the literature (cf. Burke, 1966, for a review of 14 studies). However, an examination of the research in which the theory itself, or hypotheses derived from it, was investigated, shows mixed results and conclusions. Studies in which the basic Herzberg methodology was used to collect and analyze the data typically uphold the theory, for example, Schwartz, Jenusaitis, and Stark (1963), Saleh (1964), and Myers (1964). On the other hand, studies which put job factors into a structured format (questionnaire) and use techniques for data analysis, such as factor analysis (Ewen, 1964; Friedlander, 1963), comparison of means, correlation analyses, variance analyses (Friedlander, 1964; Gordon, 1965; Halpern, 1966; Wernimont, 1966), or a scaling technique (Burke, 1966), either do not support the theory as stated, or at best give equivocal support to it. In sum, the theory has been both substantiated and refuted in part.

It seems apparent that part of the ambiguity surrounding the Herzberg theory might be attributable to methodological inconsistencies and the lack of a formal and logically consistent statement of the relations among the variables of interest. Therefore, the present study was undertaken to: (*a*) critically examine the basic Herzberg methodology, (*b*) attempt a formal statement of

[1] This study is based on a portion of a doctoral dissertation by the senior author submitted to the graduate school of The Pennsylvania State University in partial fulfillment of the requirements for the PhD.

[2] Now at the Office of Evaluation Studies, Georgia Institute of Technology, Atlanta.

the theory, and (*c*) advance and test a bipolar model of job satisfaction based on a modification of the Herzberg theory. A detailed examination of methodologies other than the one employed by Herzberg et al. (1959), and used subsequently by Schwartz et al. (1963), Saleh (1964), and Myers (1964), is beyond the scope of this paper.

Examination of the Herzberg Methodology

The basic Herzberg method of collecting and analyzing data involves providing subjects (*S*s) with two sets regarding their job experiences. First they are asked to recall a time in their job tenure when they were either (*a*) satisfied or (*b*) dissatisfied. Then *S*s are invited to discuss these events (incidents) and to specify the factors which contributed to their satisfaction-dissatisfaction experiences.

In the original study (Herzberg et al., 1959), 16 job-factor categories were identified through content analysis and collapsed into two general categories, *motivators* and *hygienes*, on the basis of logical relationships among the categories. Relationships between motivators, hygienes, and job satisfaction-dissatisfaction [3] were hypothesized by examining the frequency with which motivators were associated with job satisfaction, and hygienes with job dissatisfaction.

This basic methodology, essentially a critical incident interview, is subject to criticism for the following reasons: (*a*) it does not control either the number of incidents from a given *S*, or the number of job factors mentioned with a given incident; (*b*) it reverses the roles of the dependent (satisfaction-dissatisfaction) and independent (motivators and hygienes) variables by setting the dependent variables at one of two levels (high-low) and allowing what are conceptually the independent variables to vary as a function of the *S*'s responses; and (*c*) it does not look at the relationship of job satisfaction to motivators and hygienes across intermediate levels of job satisfaction-dissatisfaction; that is, the theory does not specify a functional relationship between the variables of interest; and (*d*) it does not provide for examination of higher degree (interaction) effects between these two classes of variables (motivators and hygienes) as they relate to satisfaction.

[3] For the remainder of this article, the term satisfaction-dissatisfaction will be employed when the concept refers to Herzberg's use of two unipolar continua; and satisfaction employed singly will refer to the traditional bipolar concept of satisfaction.

A Formal Statement of the Theory

Herzberg's major hypothesis is that satisfaction-dissatisfaction (S-DS) is a function of the classes of motivators (M) and hygienes (H). Further, Herzberg goes on from his major hypotheses ". . . that the factors leading to positive attitudes and those leading to negative attitudes would differ [Herzberg et al., 1959, p. 29]," to state at a later point:

> This hypothesis [the two-factor motivator-hygiene hypothesis] suggested that the factors involved in producing job satisfaction were separate and distinct from the factors that led to job dissatisfaction. Since separate factors needed to be considered depending on whether job satisfaction or job dissatisfaction was involved, it followed that these two feelings were not the obverse of each other. The opposite of job satisfaction would not be job dissatisfaction, but rather *no* job satisfaction; and similarly the opposite of job dissatisfaction is no job dissatisfaction—not job satisfaction. The fact that job satisfaction is made up of two unipolar traits is not a unique occurrence [Herzberg, 1964, p. 3].

The above assertions imply the following functional relationships between job satisfaction-dissatisfaction and motivators and hygienes:

$$S = f(M + e) \qquad [1]$$
$$DS = f(H + e) \qquad [2]$$
$$\rho(M,H) = 0 \qquad [3]$$

That is, satisfaction (S) is a function of motivators (M) plus other potential factors and/or error of measurement [1]; dissatisfaction (DS) is a function of hygienes (H) plus other potential factors and/or error of measurement [2]; and the correlation (ρ) between motivators and hygienes is zero [3].

Conceptually then, Herzberg posits two unipolar continua (satisfaction, dissatisfaction) which are unrelated and states that the variance in satisfaction is due to the presence and/or level of motivators, and the variance in dissatisfaction is due to the presence and/or level of hygienes. Expressions [1] through [3] above are formal statements of the relations

among the variables of interest as treated by the theory on a conceptual level.

Modification of the Theory

Rationale. The rationale for the adoption of the model of job satisfaction tested by the present research is based on a consideration of the research strategies employed by Herzberg and other investigators, and the assumption that satisfaction is more parsimoniously conceptualized as a bipolar variable.

Using the bipolar assumption, it is suggested that by examining only the extremes of the underlying continuum of satisfaction through the use of instructional sets (satisfied-dissatisfied), Herzberg and other investigators using his methodology have concluded that motivators and hygienes are disjoint, and that satisfaction and dissatisfaction should be conceptualized as two separate continua.

If the Herzbergian conclusions are formalized and translated into a bipolar model of job satisfaction, the following conditions would obtain:

For high job satisfaction (S_{++}),

$$S_{++} = f(aM + bH), \qquad [4]$$

for high job dissatisfaction (S_{-}),

$$S_{-} = f(aH + bM) \qquad [5]$$

where a is some positive coefficient and b equals zero, and M and H are as defined previously.

Expressions [4] and [5] clearly indicate what is implied by Herzberg's two continua for satisfaction, namely, the relationship of satisfaction to a given antecedent (motivator, hygiene) is not independent of the other antecedent. In other words, the effects of motivators and hygienes on satisfaction, conceptualized as a bipolar continuum, are nonadditive because the coefficient for the weight of a given antecedent is required to change as the criterion score and the other antecedent change.

In view of this rationale, it was therefore decided to adopt and test a model of job satisfaction in which job satisfaction was conceptualized as a bipolar variable, and which had the capability of being additive or nonadditive, depending on the data.

The model tested. Job satisfaction was first conceptualized as being a function of motivators, hygienes, the joint contribution of motivators and hygienes, plus other potential factors and/or errors of measurement. In view of the nonadditivity requirement, it was assumed that the function of interest is linear in the parameters (Moonan & Wolfe, 1963, p. 501), and that interaction effects exist. These considerations lead to the adoption of the following statistical model to be tested:

$$S = aM + bH + cMH + e \qquad [6]$$

where S = Satisfaction, M = Motivators, H = Hygienes, MH = Joint Contribution of motivators and hygienes, a, b, c, and e, are parameters.

Hypotheses

Three hypotheses were tested in the present research.

1. A significant proportion of the variance in job satisfaction is accounted for by motivators and hygienes.

2. The relationship of motivators and hygienes to satisfaction is linear in the parameters but nonadditive; that is, satisfaction is a joint function of motivators and hygienes.

3. A greater proportion of the variance in job satisfaction is contributed by motivators than by hygienes.

Method

The method chosen was an analogue of the direct experimental approach. In lieu of subjecting workers to different physical arrangements, the independent variables were manipulated verbally. This approach is based on the assumption that workers can accurately evaluate different conditions without being physically subjected to them at the time they are queried. The present research made this not uncommon assumption, one that is made by most attitude questionnaires, rating scales, personality inventories, etc.

A complete $3 \times 3 \times 2$ factorial design was used, with three levels each of a motivator, Factor A, and hygiene, Factor B, and two classifications of employees (professional, nonprofessional), Factor C, to test the model of job satisfaction. An independent group of 15 *Ss* was employed in each of the 18 cells of the design. The levels of the motivators and hygienes were designated (1) high, (2) medium, and (3) low.

Treatment Questionnaires

A questionnaire format was used to implement the independent conditions; that is, a section of each treatment questionnaire presented workers with

statements specifying the relevant conditions. For each of the nine treatment conditions required by the design, a separate questionnaire was designed which paired given levels of a motivator with given levels of a hygiene.

For the purposes of the present study, it was decided to select one job factor to represent the class of motivators and another to represent hygienes.

The rationale adopted for the selection was to identify the strongest or most reliable motivator and hygiene factor as denoted by its frequency of appearance in workers' stories. Thus, Achievement, which appeared 41, 55, and 44%, respectively, in three independent studies (Herzberg et al., 1959; Myers, 1964; Schwartz et al., 1963), was chosen to represent the class of motivators, and Company Policy and Administration (CPA), which appeared 31, 25, and 24%, respectively, was chosen to represent the class of hygienes.

A subset of three secondary factors relating to Achievement and three relating to CPA were selected from those listed by Herzberg et al. (1959, p. 143ff) to provide replication of the main job factors chosen. For Achievement, the three secondary factors chosen were: (1) job accomplishment, (2) job completion, and (3) solution to job problems. For CPA, the three secondary factors were: (1) organization of work, (2) benefit programs, and (3) management policies.

In addition, three other secondary factors relating to satisfaction, but not explicitly defined as a motivator or hygiene by Herzberg, were selected to be included in each of the nine questionnaires. These factors were: (1) general feelings about work, (2) pay increase, and (3) promotion. Items from these secondary factors (called General items) were written to reflect only the high level and were included in all nine questionnaires. Their purpose was to provide a common frame of reference so that the effect of the different pairings of the items could be compared to a fixed level across all questionnaires.

Item writing. The basis for the item writing was to state, as formally as possible, the three qualitative levels of the underlying continuum for each of the six secondary factors. Items were behaviorally oriented in the case of the Achievement items and referred to concrete situations in the case of the CPA items. This procedure resulted in 18 items (three levels each for the six secondary factors). A complete list of the 21 items used in constructing the nine questionnaires is given in Table A.[4]

Instructions and questionnaires formats. The instructions for the questionnaires asked the worker to think of a time in his job experience when he may have experienced the situations to be described. Then he was asked to indicate how satisfied he was with the situation at the time he was thinking of. If he had not experienced all the situations, he was asked to imagine how satisfied he would be if he had.

Each questionnaire was composed of three parts.

Part I consisted of the nine items presented singly (three Achievement, three CPA, and three General) for a given questionnaire. The level of the three Achievement and three CPA items for a given questionnaire depended on the cell it occupied in the basic 3×3 design.

The instructions for Part II, in which the items were paired, asked workers to read the pairs of items, and then, on the basis of their work experience, to indicate their degree of satisfaction with both of the situations. Part II was the key section of the questionnaire in that it assessed the satisfaction of workers with different levels of motivators and hygienes.

Part II consisted of 12 pairs of items in a single question format. The same items which appeared in Part I of a given questionnaire were used to form the pairings. The particular pairing of items for a given questionnaire depended, as for Part I in which the items were presented singly, on the cell it occupied in the basic 3×3 design. Part II of each questionnaire also contained the same three questions which were composed of three pairings of the three high General items.

A final question, in Part III, asked the worker how satisfied he generally felt about his work.

Directions were the same for all nine questionnaires. Only the pairings of the items, as required by the experimental design, were changed.

The Response Measure

Workers were asked to indicate their degree of satisfaction with the situations on a 5-point scale located below each item. The verbal descriptions of the points ranged from dissatisfied through satisfied. For the main analysis, the response measure was the sum of the nine satisfaction scores obtained from the pairings of the nine key items in Part II of the questionnaire. This score had a range from 9 to 45, with a neutral or zero point at 27.

Selection of Subjects

The *Ss* were selected from a small aerospace research and development company in Central Pennsylvania employing about 600 full-time male personnel. A roster of all full-time male personnel was obtained, and employees were then divided into a professional and nonprofessional classification by examining the educational requirements for a given job title. A given job was considered to be professional if it required a minimum of a bachelor's degree. All other jobs were classified as nonprofessional.

A total of 324 *Ss* (162 professional and 162 nonprofessional) were selected from the roster by taking approximately every other name.

[4] A complete list of the 21 items used in constructing the questionnaire and a complete sample questionnaire have been deposited with the American Documentation Institute. Order Document No. 9389 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C., 20540. Remit in advance $1.25 for microfilm or $1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

Administration of the Questionnaires

An interoffice memo, under the signature of the Director of Professional Development, was prepared. It requested the participation of the addressee in a company study on job satisfaction, explained how to complete and return the questionnaire to the office of Director of Professional Development, and guaranteed the anonymity of the respondent. This memo was sent to each *S* at his company office together with one of the nine forms of the questionnaire.

The total *N* in the sample (324) was divided into 9 groups of 36 each, corresponding to 36 replications of the 9 treatment questionnaires. The *Ss* were assigned to treatments by sending Questionnaires A through I to the first 9 *Ss* on the roster, Questionnaires A through I to the second 9 *Ss* on the roster, and continuing this process until the list was exhausted.

RESULTS

Questionnaires Received

Of the 324 questionnaires sent out, a return of 289 usable questionnaires, or 89.2% of the original sample, were received within 1 wk. The usable questionnaires were then divided into 9 groups corresponding to the 9 treatment conditions. Each group contained a minimum of 15 questionnaires for the professional and nonprofessional classification. A total of 19 questionnaires were randomly discarded from the groups which contained more than 15, leaving a final *N* of 270 (9 groups of 15 nonprofessional employees and 9 groups of 15 professional employees). The research was based on this final *N* of 270, or 83.3% of the original sample.

TABLE 1

ESTIMATED ANALYSIS OF VARIANCE RELIABILITIES FOR EACH OF THE NINE TREATMENT CONDITIONS (QUESTIONNAIRES)

Questionnaire form	Reliability
A (1,1)	.95
B (1,2)	.88
C (1,3)	.87
D (2,1)	.76
E (2,2)	.88
F (2,3)	.78
G (3,1)	.74
H (3,2)	.71
I (3,3)	.70

Note.—N = 30 *Ss* per questionnaire; number of items = 9.

Characteristics of Respondents

The mean age of the *Ss* was 30.15 yr. with a standard deviation of 6.10. In general, the respondents were quite satisfied with their work. For the nonprofessional group, 77% said they were either "somewhat satisfied" or "satisfied" with their work, while the corresponding data were 72% for the professional group.

The majority of the professional employees were engineers, and many of the department heads and branch managers also had engineering backgrounds. For the nonprofessionals, there was a wider representation of *Ss* by job titles. However, they tend to be a rather skilled group of workers who assist engineers in the development and building of specialized electronic components, or assemble limited production models of electronic equipment.

No claim is made that this group of respondents is representative of all levels of unskilled, semiskilled, and skilled workers. However, it is believed that the groups are representative of employees found in the highly competitive aerospace industry, and that the professional group represents both the managerial and technical specialist.

Reliabilities of the Response Measure

The analysis of variance model (Winer, 1962, p. 124ff) was employed to estimate the reliability of the response measure.

Shown in Table 1 are the estimated reliabilities for the nine treatment questionnaires. The numbers in parentheses beside the letter designating the questionnaire form refer to the respective levels of the motivator and hygiene that were paired in the given questionnaire. Thus, Form A (1,1) refers to Level 1 (high) of the motivator, and Level 1 (high) of the hygiene, Form B (1,2) refers to Level 1 (high) of the motivator, Level 2 (medium) of the hygiene, etc.

The estimated reliabilities of the response measure varied from .70 for the low-low pairings to .95 for the high-high pairings, with the mean of the nine reliabilities being .80. These data indicate that *Ss* are generally consistent in endorsing their degree of satisfaction elicited by the various pairings of the motivator and hygiene items.

TABLE 2

MEANS AND STANDARD DEVIATIONS OF THREE GENERAL ITEMS INCLUDED IN ALL FORMS OF THE QUESTIONNAIRES

Part I		Questionnaire form	Part II	
$\bar{X}$	SD		$\bar{X}$	SD
13.40	1.90	A (1,1)	13.73	2.21
12.57	2.72	B (1,2)	12.97	2.96
12.97	1.83	C (1,3)	13.43	2.14
12.37	2.73	D (2,1)	12.33	2.97
12.90	1.25	E (2,2)	12.70	2.78
12.43	2.46	F (2,3)	13.13	2.75
13.37	1.75	G (3,1)	13.90	1.82
13.00	1.89	H (3,2)	13.53	2.52
13.56	1.54	I (3,3)	14.07	1.70

Note.—N = 30 Ss per questionnaire; number of items = 3.

Response Bias

To check on the possibility of response bias (i.e., Ss using the particular pairing of motivator and hygiene items to anchor their satisfaction response), three General items, described in the preceding section, were included in Part I of all questionnaires, and three pairings of those items were included in Part II of all questionnaires. If no bias existed, the means of these General items could be expected to coincide in all forms of the questionnaires, whereas a bias might be expected to yield significant difference among the means. The means and standard deviations of the three General items included in Parts I and II of all questionnaires are shown in Table 2.

Examination of the means and standard deviations in Table 2 reveals that they are within sampling error for both Parts I and II of all nine questionnaire forms. The t for the differences between the highest and lowest means for Part I and Part II are both less than unity, being .389 and .508, respectively. Thus, the hypothesis of this particular response bias due to the questionnaire format and item content is thought to be highly improbable.

Main Effects

As indicated in Table 3, both of the main effects, Achievement and CPA, were highly significant, showing a reliable relationship between these factors and employee satisfaction. Factor C, employee classification, did not reach the F value at the .05 level of significance, indicating that there are no significant differences between the way professional and nonprofessional employees relate Achievement and CPA to their satisfaction.

To examine the percentage of variance independently accounted for by the main and interaction effects, a test, given by Hayes (1963, p. 406), was applied to the variance data. This test revealed that the main effects of the motivator (Achievement) accounted for over three-and-one-half times as much of the

TABLE 3

ANALYSIS OF VARIANCE SUMMARY

Source	SS	df	MS	F
Achievement (A)	15425.78	2	7712.89	317.53[a]
Company Policy (B)	4557.54	2	2278.77	93.81[a]
Employee Classification (C)	81.12	1	81.12	3.34
A × B	315.78	4	78.94	3.25[a]
A × C	98.95	2	49.47	2.04
B × C	3.95	2	1.97	.08
A × B × C	160.38	4	40.09	1.65
Error (within cell)	6120.80	252	24.29	
Total	26764.30	269		

[a] $F_{.95}(2/252) = 3.04$; $F_{.95}(4/252) = 2.42$.

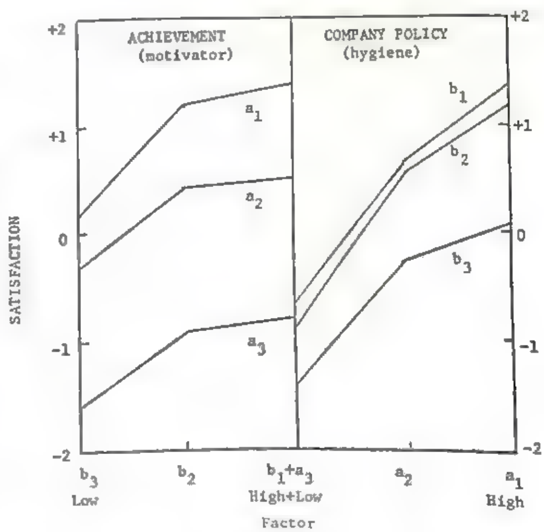


FIG. 1. Profiles of achievement at levels of company policy and company policy at levels of achievement.

variance (.574) in the criterion as did the hygiene (CPA)—.168. The interaction term, while significant, accounted for less than 1% of the variance (.008).

Interaction Effect

Profiles of the AB interactions are shown in Figure 1. The data have been transferred back to a 5-point scale of satisfaction with the midpoint being the neutral or "neither satisfied nor dissatisfied" point.

The strong effect of Achievement on satisfaction is evident from an inspection of Figure 1. Referring to the left side of the figure, even when CPA is high (b_1), low Achievement (a_3) will result in dissatisfaction. The converse is not true, however. When Achievement is high (a_1), even low CPA (b_3) will lower satisfaction to only the neutral point. Highest satisfaction is obtained when both Achievement and CPA are high, and lowest satisfaction is obtained when both are low. However, the best predictor of the criterion score is Achievement. The slopes between the three levels of Achievement (a_1, a_2, a_3) across CPA are relatively small when compared to the slopes of CPA across Achievement. One could predict a worker's satisfaction with a fair degree of accuracy given only his Achievement level.

DISCUSSION

In terms of the hypotheses tested, the results of the present research are relatively unambiguous. The two classes of antecedent variables, motivators and hygienes, accounted for nearly 75% of the variance in satisfaction scores. This finding is significant for two reasons. It indicates first that the treatment questionnaires were perceived and answered in a fairly consistent manner by the respondents. The relatively high analysis of variance estimates of the reliabilities of the treatment questionnaires are simply another indication of this fact. And second, the high percentage of variance accounted for indicates that motivators and hygienes are two classes of job-factor variables which can be meaningfully related by workers to their job satisfaction.

The second hypothesis, which predicted an interaction between the treatment variables and the criterion scores, was also confirmed, although the magnitude of the interaction was small, especially in comparison with the main treatment effects.[5] The magnitude of the interaction effect may place it in that gray area of "practical" versus "significant" differences. However, it does suggest that the independent variables (motivators and hygienes) make a joint contribution to satisfaction over and above their weighted sums; and hence the effects of motivators and hygienes are not independent of one another.

The third hypothesis, concerning the differential proportion of variance in criterion scores accounted for by motivators and hygienes, was also confirmed. Motivators accounted for nearly three-and-one-half times as much variance (.57) as did hygienes (.17), indicating that, when looking at those two classes of variables, workers view motivators as the prime source of their satisfaction. This conclusion is further strengthened by the observation that low or even moderate levels of achievement lead to reported feelings of dissatisfaction, regardless of the level of the hygiene.

These results support the structural model of job satisfaction advanced and tested by the present research. Since no significant differences were found between the professional and nonprofessional groups employed by the present research, it appears that this model is appropriate for both groups.

At the beginning of this paper, questions were raised concerning (a) the basic Herzberg methodology on which his theory of job satisfaction rests, and (b) the lack of a statement of the functional relationships between job satisfaction and motivators and hygienes. It was argued that Herzberg's conclusions about the nature of job satisfaction may be a function of his methodology and that a more rigorous statement of his theory was needed. Such a statement was attempted by the present research. It is now in order to examine

[5] It is recognized that other variables, not investigated by the present research, such as the level of overall satisfaction of workers or the type of work (skilled versus unskilled), may affect the magnitude of the interaction between motivators and hygienes. This problem is worthy of future investigation.

the results of Herzberg and prior researchers with regard to the results of the present study. To do this it will be convenient to classify prior research into two rough categories: that which lends support to the Herzberg theory, and that which tends not to support the theory.

The Herzberg Theory: Supportive Research

Three studies (Myers, 1964; Saleh, 1964; Schwartz et al., 1963) reported results which were in essential agreement with Herzberg's conclusion that motivators determine satisfaction and hygienes determine dissatisfaction. However, these studies used the basic Herzberg methodology which was criticized by the present study for a number of reasons. A fourth study by Ewen (1964) was equivocal but mostly negative. However, he provided no information regarding the instructions given to his *Ss*, hence no definite conclusions about his results can be drawn.

Friedlander's (1963, 1964) two studies provide some support for the theory. In the first (1963) he factor analyzed a 17-item questionnaire which measured the importance of the items as a source of employee satisfaction. He found three factors, two of which corresponded in part with Herzberg's concepts of motivation and hygienes, while the third factor drew from both the motivation and hygiene factors. In the second study (1964) Friedlander investigated the bipolar assumption of job satisfaction. Both correlational and variance analyses indicated that satisfaction and dissatisfaction are, for the most part, unrelated and not complementary functions. However, his method is questionable because he gave the same items to the *Ss* under two different sets and with two different emphases on the items. The first time the items were phrased positively, *Ss* were given a satisfied set, and asked to indicate the importance of the factors to their satisfaction. The second time the items were negatively paraphrased, *Ss* were given a dissatisfied set, and asked to indicate the importance of the factors to their dissatisfaction. Here one sees again that the roles of the independent and dependent variables have been reversed through instructional sets. This reversal is enough to question Friedlander's experimental logic without pointing out that it does not necessarily follow that *Ss* will ascribe the same importance to reversed items under two different instructional sets.

A final study which was interpreted by the author (Halpern, 1966) as supporting the Herzberg theory raises an interesting point. Halpern had 93 *Ss* rate aspects of their best-liked job on a 7-point graphic scale. The stimuli were four hygiene and four motivator items. In addition, he determined a rating of the *Ss*' overall job satisfaction on the same scale. He found that two items, Satisfaction with Work Itself and Opportunity for Achievement, accounted for all of the explained variance (74%), and that *Ss* were equally satisfied with both (motivator and hygiene) aspects of their job. Halpern (1966) then concluded that the basic thesis of the Herzberg theory that ". . . it is the motivators . . . that are primarily related to job satisfaction [p 200]" was supported.

However, Herzberg also stated that job satisfaction and job dissatisfaction are two separate continua which implies that those factors which determine job satisfaction (motivators) cannot be related to those factors which determine job dissatisfaction (hygienes). Since Halpern's data show substantial correlations between motivator and hygiene items, his results could also be interpreted as not supporting the theory. The point is that, with the theory as loosely formulated as it is, different interpretations are possible with the same data.

Halpern's data, however, do support one conclusion of the model of job satisfaction tested by the present research, that is, motivators and hygienes interact in determining workers' levels of job satisfaction.

The Herzberg Theory: Nonsupportive Research

The data of this study indicate that both motivators and hygienes are related to satisfaction (conceived of a bipolar variable). The results of this study further suggest (see Figure 1) that workers with strong feelings of achievement on a job will remain satisfied even though conditions surrounding the job, such as work organization, management policies, and benefit programs, are perceived

as being inadequate. On the other hand, workers who do not feel as if they are accomplishing much on the job will be dissatisfied, even though conditions surrounding the job are good. These findings are clearly at variance with Herzberg's contention that there is no interaction between motivators and hygienes.

The differential strength of the relationship between motivators, hygienes, and level of satisfaction, found in the present study, has been reported also by Halpern (1966), Gordon (1965), Friedlander (1964), Wernimont (1966), and Burke (1966). All the studies cited found that motivators are more important to satisfaction than are hygienes. Herzberg does not speak directly to this point because his methodology was not amenable to quantitative expressions of the relationships between the variables of interest.

In addition to the present study, other studies have taken issue with Herzberg's conclusion that specific job factors affect satisfaction in only one direction. Gordon (1965), Wernimont (1966), and Burke (1966) concluded that motivators and hygienes are not unidirectional in their effects. And while Friedlander (1964) could find little evidence for the bipolarity of job satisfaction, neither did his results suggest that it is motivators which determine satisfaction and hygienes which determine dissatisfaction.

In summary, it is seen that those studies which directly support the Herzberg theory have employed a questionable methodology which was criticized by the present research and others (Brayfield, 1960; Vroom, 1964). On the other hand, the present research and other studies which did not directly support the Herzberg theory have controlled for response variability by using a structured format, looked at the relationship of motivators and hygienes to more than two levels of satisfaction, and have either implicitly or explicitly conceptualized satisfaction as a bipolar variable.

Therefore, the evidence points to the possibility that the main conclusions of the Herzberg theory (i.e., satisfaction is determined by motivators; dissatisfaction is determined by hygienes; satisfaction and dissatisfaction are two unipolar continua; and the correlation between motivators and hygienes is zero) are a function of the methodology evolved by Herzberg.

Brayfield (1960) has suggested this possibility, and Kahn (1961) and Vroom (1964) have suggested that defensive behaviors and displacement could account for the Herzberg findings. In addition, Vroom (1964) and Vroom and Maier (1961) have cautioned against the risk in inferring the actual causes of satisfaction and dissatisfaction from description of events by individuals.

To fully resolve the issue of the polarity and/or dimensionality of job satisfaction and the relationship of the variables of interest within the Herzberg theory, further research is necessary. It is believed that the evidence from the present study as well as that from other studies cited, especially Burke (1966), is strong enough to cast doubt on Herzberg's double-dichotomy conception of job satisfaction, that is, motivators versus hygienes, satisfaction versus dissatisfaction. It may be that the model of job satisfaction suggested by this research, using only one motivator and hygiene, is biased, and an examination of the remaining or other components of job satisfaction would alter this model. A recent study by Hulin and Smith (1965) is relevant. They attempted to predict five components of job satisfaction (satisfaction with work, pay, promotions, co-workers, and supervision) from a linear combination of six independent variables, but found only satisfaction with work and pay to be significantly related to a set of independent variables. Their data suggested that the five aspects of job satisfaction under investigation were different from each other in terms of functional relationships and predictability; and further, that a linear model of job satisfaction could best explain the findings of their study.

Therefore, it is evident that a more definitive study of the Herzberg theory should examine all of the components of satisfaction, the model which best fits the data, as well as the strength of association between the independent and dependent variables.

Generality of Present Research

Although the relationship between motivators, hygienes, and satisfaction was found

to be the same for both professional and nonprofessional employees, the generality of the present research is limited by four conditions:

1. The data are based on employees from one company.

2. The employees were, in the main, either skilled or professional.

3. For the most part, *Ss* were quite satisfied with their work.

4. All *Ss* were male.

Conclusions

The following conclusions are based on the results of the research and their interpretation.

1. The class of job factors, known as motivators and hygienes in the Herzberg theory of worker satisfaction, can be used to describe most of the variance in job satisfaction.

(*a*) However, the disjoint relationship between motivators and hygienes, as postulated by the Herzberg theory, was not found in the present research.

2. Motivators and hygienes appear to be related to job satisfaction in a nonadditive fashion. This means that a given level of job satisfaction cannot be predicted from a simple weighted sum of the levels of motivators and hygienes.

3. Motivators are more important to job satisfaction than are hygienes; this importance being on the order of 3:1.

4. Pending further research, Herzberg's conception of job satisfaction as being comprised of two unipolar continua should be reevaluated.

REFERENCES

Brayfield, A. H. Review of Herzberg, F. et al., *The motivation to work*. *Personnel Psychology*, 1960, **13**, 101–102.

Burke, R. J. Are Herzberg's motivators and hygienes unidimensional? *Journal of Applied Psychology*, 1966, **50**, 317–321.

Ewen, R. B. Some determinants of job satisfaction: A study of the generality of the Herzberg theory. *Journal of Applied Psychology*, 1964, **48**, 161–163.

Friedlander, F. Underlying sources of job satisfaction. *Journal of Applied Psychology*, 1963, **47**, 246–250.

Friedlander, F. Job characteristics as satisfiers and dissatisfiers. *Journal of Applied Psychology*, 1964, **48**, 388–392.

Gordon, G. G. The relationship of "satisfiers" and dissatisfiers to productivity, turnover, and morale. *American Psychologist*, 1965, **20**, 499. (Abstract)

Halpern, G. Relative contributions of motivator and hygiene factors to overall job satisfaction. *Journal of Applied Psychology*, 1966, **50**, 198–200.

Hayes, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.

Herzberg, F. New approaches in management organization and job design—I. *Industrial Medicine and Surgery*, 1962, **31**, 477–481.

Herzberg, F. The motivation-hygiene concept and problems of manpower. *Personnel Administration*, 1964, **27**, 3–7.

Herzberg, F., & Hamlin, R. M. Motivation-hygiene concept and mental health. *Mental Hygiene*, 1961, **45**, 394–401.

Herzberg, F., & Hamlin, R. M. The motivation-hygiene concept and psychotherapy. *Mental Hygiene*, 1963, **47**, 384–397.

Herzberg, F., Mausner, B., & Snyderman, B. B. *The motivation to work*. New York: Wiley, 1959.

Hulin, C. L., & Smith, P. C. A linear model of job satisfaction. *Journal of Applied Psychology*, 1965, **49**, 209–216.

Kahn, R. L. Review of Herzberg, F. et al., *The motivation to work*. *Contemporary Psychology*, 1961, **6**, 9–10.

Moonan, W. J., & Wolfe, J. H. Regression and correlation. *Review of Educational Research*, 1963, **33**, 501–509.

Myers, M. S. Who are your motivated workers? *Harvard Business Review*, 1964, **42**, 73–88.

Saleh, S. D. A study of attitude change in the preretirement period. *Journal of Applied Psychology*, 1964, **48**, 310–312.

Schwartz, M. M., Jenusaitis, E., & Stark, H. Motivational factors among supervisors in the utility industry. *Personnel Psychology*, 1963, **16**, 45–53.

Vroom, V. H. *Work and motivation*. New York: Wiley, 1964.

Vroom, V. H., & Maier, N. R. F. Industrial social psychology. In *Annual Review of Psychology*, 1961, 413–446. Palo Alto, Calif.: Annual Reviews, Inc.

Wernimont, P. F. Intrinsic and extrinsic factors in job satisfaction. *Journal of Applied Psychology*, 1966, **50**, 41–50.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received August 8, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 4, 340-345

DYNAMIC CENTRAL SCOTOMETRY[1]

T. SHIPLEY, R. WAYNE JONES, AND A. FRY

Department of Ophthalmology, University of Miami

Using radioactivated light sources and 2 levels of preadaptation, with 7 Ss, the size of the central scotoma is measured as a function of time in the dark. The recovery time to initial sighting is shown to be a function of the level of preadaptation, but the subsequent rate of scotoma collapse is not. Thus this rate of decrease in size is a reliable psychophysical index of vision in normal observers. The possible applicability of this measure to questions of vision screening is noted.

Whenever the eye is adapted to a bright light its threshold is raised. The time it takes for the eye to recover its sensitivity after the adapting light is turned off is called the "recovery time" (e.g., Chilaris, 1962; Fry & Miller, 1964). The measurement of this time, for either fixed or constantly decreasing intensities, is the now classical procedure for determining normative dark-adaptation curves and related threshold functions (e.g., Hecht, Haig, & Wald, 1935; Hecht, Haig, & Chase, 1941). In various forms these procedures are widely applied to clinical (e.g., Forsius, Krause, & Eriksson, 1964a, 1964b) and to engineering (e.g., Severin, Newton, & Culver, 1962) problems. Terms such as dazzle recovery, photostress, and flash blindness recovery have been used to describe this event.

Instead of measuring the time for a given retinal region to recover, it is also possible to measure the size of the retinal region which, at a given time and for a given intensity, has *not yet recovered*. Thus, one measures the size of the retinal region which is insensitive, that is, is blind or scotomatous, to the stimulus that is presented. One speaks of an after-image scotoma (Locke, 1963) or of visual field measures in darkness (e.g., François & Verriest, 1953a, 1955, 1957). For very dim intensities, the central regions of the retina remain insensitive even after an indefinite period of rest in darkness. One then speaks of the "absolute" central scotoma (e.g., François & Verriest, 1953b). The measurement of this blind region is referred to as central scotometry. This scotoma is a normal physiological event which can readily be measured with large stimuli, say of 1° or greater. It exists because the peripheral threshold up to about 17° (as mediated by many rod cells converging upon one ganglion cell) is significantly better (i.e., lower) than the central threshold (as mediated by only one or only a few cone cells converging upon one ganglion cell)—Hecht et al., 1935.[2] This scotoma may be particularly large in blue light.

Since different retinal regions may take different times to reach a given sensitivity level, it is possible to use a fixed low-intensity light source and to measure the time in the dark after light adaptation that it takes for this light source to be seen in these different regions. The apparatus should be arranged in such a way that the light source may be moved continuously in the visual field, or moved at will. Thus a dynamic measure may be obtained of the regions which, at various times, are insensitive to this light. This is a dynamic plot of changing field size, or relative scotoma versus time in the dark: thus—dynamic scotometry.

THE PRESENT EXPERIMENT

The present experiment measures the size of the relative central scotoma (after a fixed light adaptation and to a target of fixed size and intensity) in this dynamic fashion. Its purpose is twofold: first, to determine the basic normative shape of the function "size of central scotoma against time in the dark," and second, to attempt to relate this function to individual differences in the visual performance

[1] Supported by contract #DA-49-193-MD-2344 from the Office of the United States Surgeon General.

[2] For much smaller stimuli (e.g., 2.7 min. of arc) the situation is reversed and a *peripheral* scotoma may be found (cf. Arden & Weale, 1954). This is also true for large targets (1°) in the far periphery—beyond 10°–20°—where the sensitivity curve reverses and falls (cf. Sloan, 1947).

of normal observers. It is assumed, for example, that for any given stimulus condition the observer (*O*) with the smallest central scotoma will naturally have the highest potential for good visual efficiency. To this latter end, the apparatus must be readily portable and compatible in its use with typical clinical procedures.

APPARATUS AND GENERAL PROCEDURE

The apparatus is shown schematically in Figure 1. The light targets (T_l, T_r) are two small 10 min. of arc discs of radioactivated carbon-14 phosphors (United States Radium Corporation) with luminances matched at about 0.6 μL. (c. 1.9×10^{-9} cd/m²). The predominately blue-green emission curve of the phosphor is also shown in Figure 1. (The amount of radioactivity involved is negligible.) The intensity of these sources will remain constant for the life of the instrument. However, if they are inadvertently exposed to light it is necessary to wait a few hours before resuming testing. For this reason, each disc is carefully mounted in a holder with a light-tight cap which remains closed until all room lights are extinguished.

In Figure 1, the subject (*S*) is seen in front of the scotometer and the experimenter (*E*) behind. The *S* is tested monocularly. The preliminary light adaptation is given by a hemispherical light source. Any wide-field light source of relatively homogeneous white light will do. The *S* must be carefully instructed to keep his eyes open and to look straight ahead during the preadaptation period to assure a constant state of light adaptation.

After this, *S* pivots on a swivel chair and moves into the headrest. He then fixates the small red fixation light (F). Since it takes at least 30 sec. for *S* to see this light after the light adaptation has been completed, this maneuver in no way disrupts the continuity of the experiment provided only that the room is relatively light tight.

After the light-adaptation light is turned off and *S* is fixating, *E* monitors this fixation by raising and lowering the brightness of the red fixation light and requesting *S* to describe what is happening. When *E* is certain that *S* is fixating properly, he lowers the intensity of the fixation light until it is just above threshold, and henceforth keeps it at this level. The *E* now opens the cap covering the test stimulus and exposes it to *S*. The *S* is instructed to indicate (by tapping) when he sees this light. He is constantly cautioned not to move his eyes from the fixation point.

On the basis of the classical findings of Hecht et al. (1935), although extrapolating from their 2° stimulus to a 10 min. of arc one, the prediction was made that the target would be seen sooner in the periphery than near the fovea. Therefore, the disc was first presented in the farthest peripheral position permitted by the instrument, and worked inward from there. The single target, T_l or T_r as the case may be, was presented successively in each of the four cardinal retinal directions. The arm (A) carrying the target was pivoted by *E* into these positions.

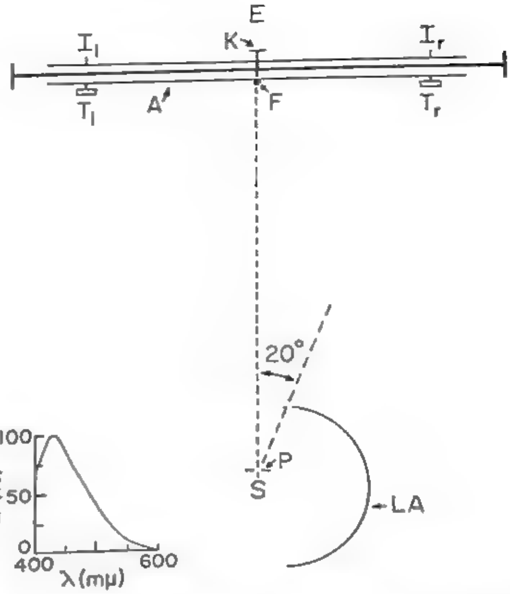


FIG. 1. Schematic drawing of the apparatus. (LA, light-adaptation dome; F, fixation point; A, rotating arm; T_l, T_r, targets; I_l, I_r, degrees indicators; K, control knob; P, artificial pupil; S, subject; E, experimenter.)

The sequence of these directions was systematically counterbalanced between *Ss* and experimental sessions to control for bias. The *S* was told in which direction to expect the light to appear. As soon as *S* saw the light and began tapping, *E* recorded the time and the target position, as given by the degrees indicator (I_l or I_r). The *E* then moved the target inward, by means of the control knob (K), to where *S* ceased to see it. Then *S* again waited until it was seen—this process was iterative until the absolute central scotoma was reached.

Experiment 1

The testing of untrained *Ss* was preceded by studies on trained *Os* in order that the principles involved might first be elucidated. For this experiment the authors served as *Ss*. As noted, the disc was 10 min. of arc to match the common clinical perimetric target (3/1000). The pupils were dilated with mydriacyl. The *S* first dark-adapted for 10 min. Then followed a light adaptation of 2 min. to 41.4 cd/m² of white light (obtained with a fluorescent lamp adaptometer of the authors' design, placed as in Figure 1). A 5-mm. artificial pupil was worn about 15 mm. before the cornea during the scotometry measurements (not during light adaptation). Under these conditions, the maximum possible peripherality of the target was 16.51°.

Since the eye is held still throughout this experiment and since the target is simultaneously both far in the periphery and below cone threshold, it frequently fades in and out in accordance with the Troxler effect. All threshold perimetric studies

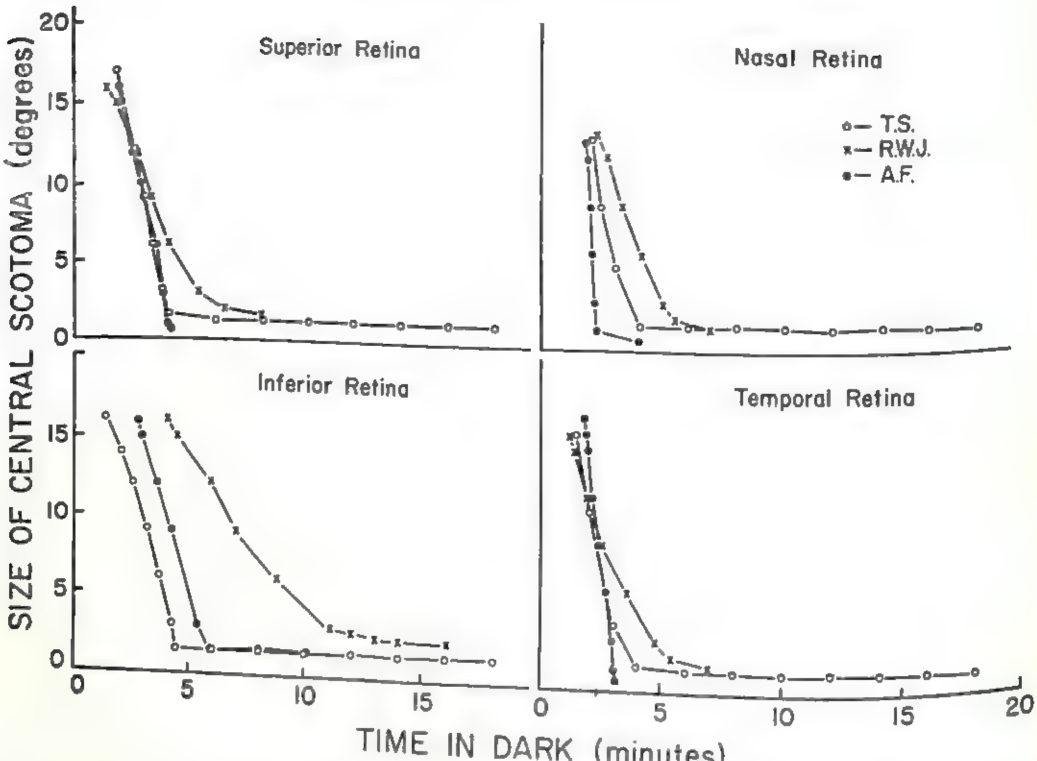


Fig. 2. The size of the central scotoma as a function of time in the dark for a blue-green target of 10 min. of arc and 0.6 μL. ($N = 3$ trained observers.)

will be confounded by this effect, and it must be said that the results of this study were derived in the face of this randomness. The attempt was made to counter this by oscillating the target back and forth over a few degrees whenever it remained invisible for what seemed to be an inordinately long period of time.[3] This often served to bring the target into view. Other than this, the target was moved at a consistent and steady rate from *O* to *O*.

All three *O*s noticed that the target became brightest at about 12° and stayed this way until about 9°–8°, after which time it again became significantly dimmer. This agrees in general with the comments of many authors (cf. François & Verriest, 1953b) concerning an annulus of maximum absolute sensitivity at about 10°–14° off the fovea, but it is not in accord with the position of maximum rod density at 17°, nor with Sloan's (1947) finding of a minimum point at 17°. Quite probably, considerations of field size (Sloan's 1° versus the 10 min. of arc of this study) and of ganglion cell density are involved here as well as of rod density. Moreover, visual acuity is very poor beyond 12°, and the major factor operating there must be some crude dyscritic threshold mechanism based upon rod convergence. Here the target appears only as an amorphous region of luminosity. Inside of say 5°, visual acuity improves and *S* becomes aware of the emergence of epicritic vision as the target takes on a visibly circular shape.

Despite these difficulties, the results for all retinal directions and for all *O*s converge rather nicely. These are shown in Figure 2, where we have plotted the size of the central scotoma (as ordinate) against time in the dark (as abscissa). Each curve is the best fitting curve of four runs. Observer A. F. (O_{AF}) drops rapidly and has no measurable residual scotoma. (N.B.: the instrument could not measure a scotoma under 10 min. of arc in size or within 0.7° of the fovea.) Observer R. W. J. (O_{RWJ}) drops significantly more slowly in all four directions and has a scotoma of 2.4° along the inferior retinal axis. Observer T. S. (O_{TS}) falls at intermediate rates and has a very slight scotoma along the nasal axis.

These data indicate that, for a 10 min. of arc target, the retinal periphery adapts only very slightly faster than does the perimacular region. Second, following the previous argument, they show that it is possible to rank these *O*s in terms of their potential skill in night visual tasks: O_{AF}, O_{TS}, and O_{RWJ}, in that order. While these *O*s have identical normal epicritic vision under photopic conditions (20/20 acuity), their *foveal* dark-adaptation curves also show this same general rank order though to a lesser degree.

[3] One could, of course, use a shutter to flicker the light, a procedure known to reduce the Troxler effect, but this adds both a mechanical inconvenience and a new stimulus variable to an otherwise uncomplicated presentation.

Experiment 2

This experiment proceeded in principle as the first, except for the taking of standard Goldmann dark-adaptation curves (3 min. light adaptation to 1,017 cd/m², 5° tungsten white target, 11° above the fovea, natural pupils). The *S* rested in relative darkness for 10 min., during which time he was instructed in both test procedures, dark adaptation and scotometry. Each *S* came for two 90-min. sessions. In each session one standard dark-adaptation curve (30 min.) was taken, and four scotometry measurements (about 10–15 min. each) were taken, one in each of the four cardinal retinal directions. Rest periods were interspersed when necessary. The experiment was particularly demanding on naïve observers, but the results suggest that measurements in only one direction (probably temporal) would probably suffice for mass testing. Scotometry light adaptation was 2 min. to 1,017 cd/m². The results for two of the authors and five naïve *Ss* are given in Figure 4. Since this figure presents a transformation of the raw data (unlike Figure 2), it is necessary to describe how it was plotted.

This may be done by referring to Figure 3 where the raw data were plotted, as in Figure 2, for two *Os*: one *O* (S. S.—O_{SS}) giving unreliable data and the other, O_{RWJ}, giving data of about average reliability. By reliability is meant the repeatability of the length of time it took, along a given direction, to first see the target. Although this time differs significantly for O_{RWJ} along the various meridians, it repeats reasonably well for a given meridian. For O_{SS}, on the other hand, one sees that the time of first sighting varies widely from session to session even along the same meridian.

Two points emerge: the first is that, unlike Experiment 1, where the lapsed time to first sighting

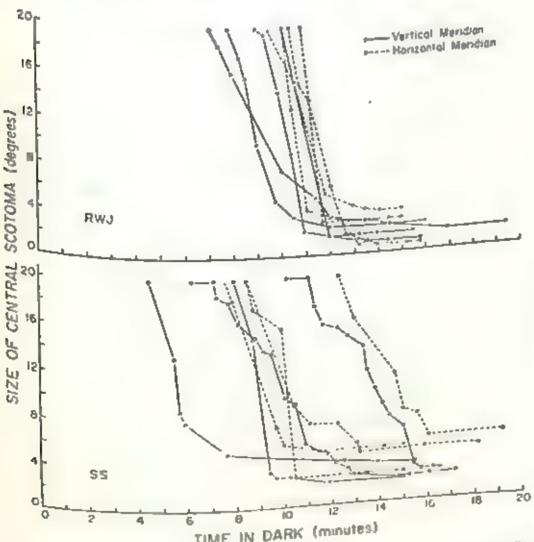


FIG. 3. An illustration of the method used to plot the data: (*a*) raw data from trained O_{RWJ}; (*b*) raw data from naïve O_{SS}.

TABLE 1

LAPSED TIME FOR OBSERVERS IN EXPERIMENT 2

Observer	Lapsed time[a]	*SD*
O_{JB}	5.22	1.22
O_{WF}	7.90	0.99
O_{SS}	8.30	2.79
O_{MB}	8.37	1.68
O_{RK}	8.46	1.02
O_{RWJ}	9.30	1.63
O_{AF}	9.35	0.49

[a] In minutes.

was only about 1–2 min., the lapsed times here are large, running up to as much as 12 min. The mean lapsed time for all *Os* and all meridians is 8.13 min. with a standard deviation of 1.41. The lapsed times for each observer are given in Table 1. This long initial delay must be due to the much higher level of preadaptation obtained in this experiment in comparison to that of Experiment 1: that is, 2 min. to 1,017 cd/m² instead of 2 min. to 41.4 cd/m². Physiologically, this is probably because the rods were saturated in Experiment 2 and not in Experiment 1.

One could thus probably derive a parameter pertinent to night vision by a systematic variation of the light-adaptation conditions. This parameter might be lapsed (or recovery) time weighted, somehow, by its own reliability. Perhaps it is the failure to take into consideration the reliability of the recovery time that has, in the past, led most workers to reject lapsed time as an index of night vision. The variations in this time within and between *Ss* may be due to the operation of factors which some *Ss* cannot consistently control, such as squinting during light adaptation or improper fixation during dark adaptation, which might also hinder them in other visual tasks of high demand.

Second, and more important, the point emerges, even from the data of O_{SS} and despite the variations in lapsed times, that the slopes of the curves, that is, the rates at which the scotoma collapses or shrinks with time (thus degrees/minute), remain fairly constant. For example, the slopes for O_{RWJ} are the same in Figure 2 and in Figure 3 despite the great difference in light adaptation. And they are approximately the same as those for O_{SS} in Figure 3 despite her poor reliability. Since it is this rate in which there is interest, the data have been replotted to permit the ready comparison of slopes. To this end, the authors took the time of first sighting the target at 14° as time zero and expressed all other times as minutes from this. Each curve in Figure 4 is thus a raw-data curve slid along the abscissa to the left by the time it took to first see the target at 14°. Figure 4 also includes the standard Goldmann dark-adaptation curves for comparison with the scotometric curves.

It appears that all *O*s display the same basic function. But O_{SS} stands out because of her large absolute scotoma (in one instance as large as 5°); and *O* R. K. (O_{RK}) stands out because of his smaller slopes during the critical early minutes; and perhaps O_{AF} also stands out because of her very large early slopes.

Discussion

Individual differences are displayed by this procedure in at least four ways: time to initial sighting (Table 1, column 2); variance in time to initial sighting (Table 1, column 3); slope or rate of collapse of central scotoma (Figure 4); and size of final absolute central scotoma (Figure 4). Such individual differences do *not* appear between these *O*s tested with standard photopic acuity measures (20/20) or with standard Goldmann dark-adaptation curves.

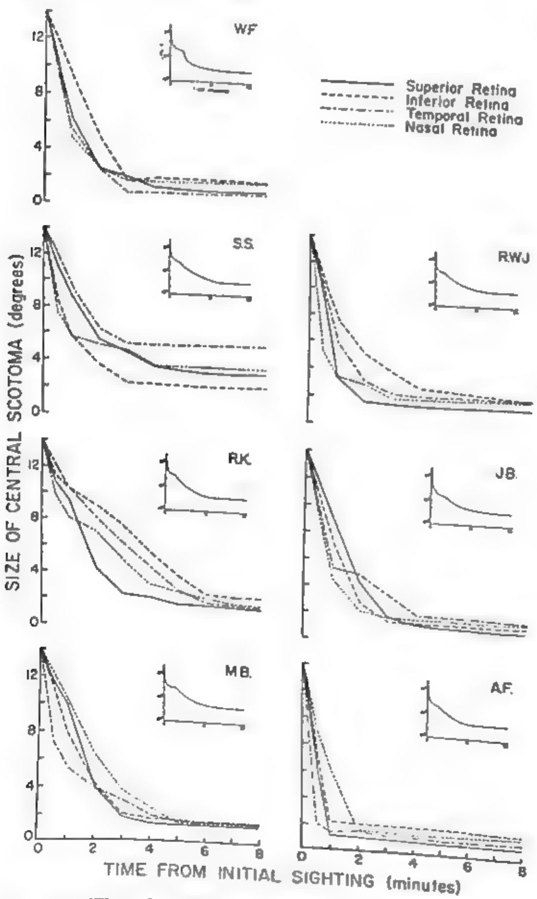


Fig. 4. The size of the central scotoma as a function of time in the dark for a blue-green target of 10 min. of arc and 0.6 μL. ($N = 7$, five observers are naïve.)

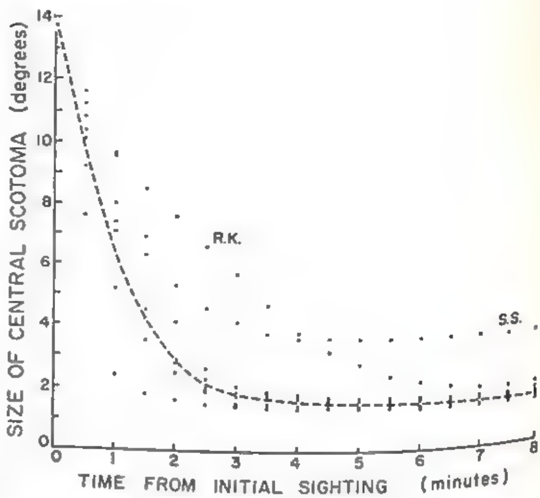


Fig. 5. Final composite curve.

Although it cannot be asserted that these differences will be related to visual performance under high demand situations (e.g., night vision), it would seem that they must be. It has already been suggested that *O* with the largest central scotoma, whether relative (O_{RK}) or absolute (O_{SS}), will have the least potential for good visual performance at low intensity levels. Similarly, one expects the unreliable observer (say *O* M. B.—O_{MB}) to be worse than the reliable one (say O_{AF}). The procedure of dynamic scotometry might thus be further explored and attempts made to correlate the results with individual differences obtained in other night-vision tasks.

Most important, however, is the conclusion which emerges from the data of Figure 4, that the method is sufficiently reliable so as to allow the authors to display a basic physiological function. Despite individual differences among the five normal *O*s, the reliability of the slopes (excluding O_{SS} and O_{RK} as has already been noted) is quite as good as that obtained for other comparable psychophysical functions.

On this basis, the mean data for all *O*s have been combined into a scattergram, and a best-fitting monotonic curve has been drawn through them, as shown in Figure 5. The dashed line is thus the mean composite curve (from Figure 4) for five *O*s and for four retinal directions and two trials. This curve may probably be assumed to hold in a normative sense for the whole retina. It supports the

concept of the radial symmetry of the retina, at least in terms of the organization of rod-ganglion cell interactions. And it shows (for the 0.6 μL.–10 min. of arc blue-green target) that the same relatively low level of rod sensitivity is attained within 3 min. over the whole important central rod retinal field of 14°–2°.[4]

This fast slope implies that these thresholds involve only a relatively *un*converged rod-ganglion cell population and, because of this, show little if any advantage in rod as opposed to cone vision.

REFERENCES

ARDEN, G. B., & WEALE, R. A. Nervous mechanisms and dark-adaptation. *Journal of Physiology,* 1954, **25,** 417–426.

CHILARIS, G. Recovery time after macular illumination. *American Journal of Ophthalmology,* 1962, **53,** 311–314.

FORSIUS, H., KRAUSE, U., & ERIKSSON, A. Dazzling test in central serous retinopathy. *Acta Ophthalmologica,* 1964, **41,** 25–32. (a)

FORSIUS, H., KRAUSE, U., & ERIKSSON, A. The dazzling test in diseases of the retina. *Acta Ophthalmologica,* 1964, **41,** 55–63. (b)

[4] This is contrary to our predictions from the 2° study of Hecht, Haig, and Wald, 1935, and its significance is discussed in a separate paper by Shipley, Jones, and Fry, in press. Scotometry measurements with a 2° target indicate an even faster drop: taking only approximately 30 sec. to go from 14° down to 2°.

FRANÇOIS, J., & VERRIEST, G. Campimetrie à l'obscur (note préliminaire). *Annales d'Oculistique,* 1953, **186,** 1–32. (a)

FRANÇOIS, J., & VERRIEST, G. Nouvelle contribution à l'étude du scotome central à l'obscur chez les sujets normaux. *Bulletin de la Société Belge d'Ophtalmologie,* 1953, **104,** 1–31. (b)

FRANÇOIS, J., & VERRIEST, G. La campimetrie à l'obscur en neuro-ophtalmologie. *Annales d'Oculistique,* 1955, **188,** 1–65.

FRANÇOIS, J., & VERRIEST, G. Une nouvelle instrumentation pour l'étude du champ visual scotopique. *Ophthalmologica,* 1957, **133,** 45–52.

FRY, G. A., & MILLER, N. D. Visual recovery from brief exposures to very high luminance levels. *School of Aerospace Medicine—Technical Documentary Report—64—36,* 1964, 1–26.

HECHT, S., HAIG, C., & CHASE, A. M. The influence of light adaptation on subsequent dark adaptation of the eye. *Journal of General Physiology,* 1941, **24,** 831–849.

HECHT, S., HAIG, C., & WALD, G. The dark-adaptation of retinal fields of different size and location. *Journal of General Physiology,* 1935, **19,** 321–337.

LOCKE, J. C. After-image scotometry: A method of functional examination of the eye. *Transactions of the American Ophthalmological Society,* 1963, **61,** 682–769.

SEVERIN, S. L., NEWTON, N. L., & CULVER, J. F. A new approach to the study of flash blindness. *Archives of Ophthalmology,* 1962, **67,** 578–582.

SHIPLEY, T., JONES, R. W., & FRY, A. Dark adaptation: Some comments on retinal position and target size. *Journal of Physiology,* in press.

SLOAN, L. L. Rate of dark adaptation and regional threshold gradient of the dark-adapted eye. *American Journal of Ophthalmology,* 1947, **30,** 705–720

(Received August 22, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 4, 346–351

FEEDBACK IN A COMPLEX MULTIMAN-MACHINE SYSTEM[1]

IRWIN L. GOLDSTEIN,[2] JACK F. SOUTHARD,[3] AND DAVID A. SCHUM[4]

Ohio State University

Human decision makers provided hypotheses and made diagnoses, in the form of conditional probability judgments, to account for the occurrence of certain critical events in a simulated hostile environment. The decision makers' probabilistic estimates were compared with similar estimates provided by a Bayesian model for several levels of percentage of knowledge of results (KR; 0%, 33%, 67%, and 100%) and 2 levels of specificity of KR (access or no access to model estimates). The data indicated that there were no significant differences in the probabilistic estimates provided for 33%, 67%, and 100% KR but that all 3 were superior to 0% KR. The human decision makers with access to Bayesian model estimates as feedback were not able to improve their judgments significantly even though the model-generated solutions were significantly superior to human estimates at all KR levels above 0%.

A series of experiments (Schum, Goldstein, & Southard, 1966) have investigated variables affecting the decision maker's ability to select or formulate hypotheses to explain the occurrence of a certain set of observed data. These investigations involved diagnosis of threats in a man-machine systems format. The decision maker provided hypotheses and made diagnoses, in the form of conditional probability estimates, to account for the occurrence of certain critical events in a simulated hostile or potentially hostile environment. It happens that there is a formal quantitative statement, Dodson's (1961) revision of Bayes' theorem for multiple-data classes, which describes how one ought to revise his opinions about the probability of some hypothesis in the light of new data or experience. In studies investigating reduced fidelity of input information (Schum, Goldstein, & Southard, 1965a) and restriction of data in terms of time available to watch the development of critical events (Schum, Goldstein, & Southard, 1965b), humans were, on occasion, able to provide higher probability estimates than the Bayesian model in the correct-hypothesis category. However, human decision makers were always inferior to the Bayesian model in terms of identifying the correct hypothesis category with first-choice estimates. Briggs and Schum (1965) have pointed out that interpretations regarding these investigations of decision making have been limited in that those subjects (*S*s) providing posterior probability estimates were always provided with complete knowledge of results about the true state of the world. The present experiment was concerned with the investigation of the effects of this variable.

In general, the term "knowledge of results" (KR) refers to information given to an individual or group of individuals about the outcome of a response or group of responses. Enough research has been conducted to warrant the conclusion that KR is of great value in a large variety of tasks including performance on an ergograph (Arps, 1917), grading handwriting (Gilliland, 1925), and performance on a complex arm-leg coordination task in industry (Lindahl, 1945). It is difficult, however, in a complex task involving human *S*s to state the effects of systematic variations in percentage of KR or specificity of KR. Studies varying percentage of KR are few in number and suggest, according to Kimble (1961), that performance ". . . is usually, but not always somewhat poorer . . . [p. 161]" for partial than for continuous conditions. Both Kimble (1961) and Michels

[1] The research reported in this paper was carried out at the Human Performance Center and was sponsored by the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, under Contract No. AF 33(615)-2248 with the Ohio State University Research Foundation. Further reproduction is authorized to satisfy the needs of the United States Government.

[2] Now at the University of Maryland.

[3] Now at Miami University (Ohio).

[4] Now at Rice University.

(1955) report a marked reduction in learning when the level of reinforced trials falls below 20% or 25%. In a complex man-machine system format it is difficult to predict how much feedback is necessary to sustain performance, but it does appear that 100% feedback may be superfluous.

The other KR variable investigated in this study was specificity of KR. This typically refers to the precision with which error is indicated in a single dimension of performance. Increased specificity involves two components: one, a directional component which states a relationship between *S*'s response and the desired correct response; and the other, the number of discriminable response categories available to *S* (Annett, 1961). It appears that increases in specificity which involve the second component do not always result in superior performance (Eriksen, 1958; Green, Zimiles, & Spragg, 1955) and may, in fact, even degrade performance through the confusion generated by such unnecessary detail (Crafts & Gilbert, 1935). It seems reasonable to state that there is little benefit from using a wider range of categories in KR than *S* has available in his potential response repertoire (Annett, 1961).

METHOD

General task. The simulation facility described in this article is applicable to a wide variety of situations involving multiple levels of information-processing and decision-making functions. The particular format chosen for this investigation was that of an intelligence agency at some unspecified but presumably high decision level.

In the present experiment a computer facility provided the means for generating a complex real-time stimulus environment simulating the movements of the surface and air forces of a hypothetical adversary called "aggressor." A team of *S*s designated as intelligence staff officers (ISOs) attempted to identify the events taking place in aggressor's territory through the use of simulated overflights. Their information was communicated through closed-circuit TV to another group of *S*s designated as decision makers (DMs) whose job it was to evaluate the threat posed by orderly buildups of aggressor forces appearing in the environment.

Subjects. All *S*s were college upperclassmen who were paid on an hourly basis. Every *S* received initial training consisting of 60–114 hr. of lectures, problems, and on-the-job training as well as further instructions specific to the experiment. All decision makers had served at least 6 mo. in previous man-machine system experiments.

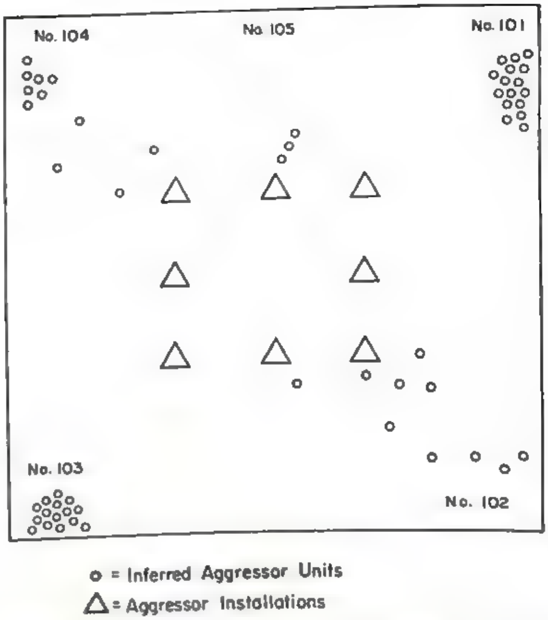


FIG. 1. A schematic diagram of aggressor territory (highly simplified).

Stimulus environment. Aggressor territory was defined as a square area 1,024 mi. on each side. Within this territory, in a time-dependent fashion, there occurred orderly buildups of aggressor forces which were called *developmental groupings.* In the present experiment there could be as many as 25 individual developmental groupings in various stages of buildup in aggressor territory at any one time. Figure 1, a highly simplified abstraction of the basic environment, illustrates this buildup process. The large triangles in the figure represent certain fixed installations such as forts, supply depots, and airfields. The clusters of small circles represent units of aggressor's surface and air forces. Each cluster represents developmental groupings which are in various stages of buildup. A developmental grouping reached the terminal stage of buildup when all of its associated units had reached their final location along one of the four borders. When this occurred, the developmental grouping remained there for a set period of time and then terminated by disappearing from the environment. System load was defined as the number of developmental groupings terminating in an experimental session. In the present experiment, six groupings terminated every session.

Information-processing task (ISO). All input information occurred only at the request of one of the four ISOs, each of whom was responsible for a different major class of information (main attack, combat support, logistics, or order of battle). A chief of staff initiated the starting procedure by choosing an area of aggressor territory for a simulated reconnaissance overflight. The computer then made available simulated photographic, radar, and

TABLE 1

PERCENTAGE OF JUDGMENTS RECEIVING FEEDBACK

Basis of feedback	0%	33%	67%	100%
I	E,C	A,F	D,G	H,B
II	E,C	A,F	D,G	H,B

Note.—I = DM's own judgment of P(H|D). II = DM's and Bayes' judgment of P(H|D). A–H = Hypothesis categories.

infrared sensor data in verbal and numerical form. The next stage was initiated by the ISOs who interrogated computer storage through their display consoles. The information thus obtained provided the ISO with descriptions of the type, number, activity, and location of collections of mobile weapons, vehicles, and aircraft. Using reference tables which related numbers and types of elements to aggressor units, the ISOs attempted to infer the existence of the various aggressor units in a developmental grouping and the spatial and temporal arrangements of these inferred units. This was a complex time-oriented process which was complicated further by the extent of the environment.

The ISO obtained as much information as possible about a particular deployment and then attached probability estimates to the various levels of each predictor. There were 25 such predictors (attribute classes), each associated with a varying number of possible states. For example, in Attribute Class 10, a given deployment might have had zero, one, two, or three tactical air support squadrons (i.e., this predictor has four possible states). The probability estimates required of ISO were intended to reflect his confidence in the number of tactical air squadrons actually present in the deployment. Thus, it was necessary for him to assign a probability value from 0 to 1 to each of the four states, and for these four values to sum to 1.00. He could be confident, for example, at .70 that there was one tactical air squadron (State 3), but he could allow for underestimation and overestimation and place probability estimates of .15 in State 1 and State 3.

Since this was a time-oriented environment, information obsolesced rapidly and ISOs were required to interrogate the environment many times. Final probability estimates made during the terminal buildup were sent by closed-circuit TV to the decision makers.

Decision-making task (*DM*). The decision makers received the probability information from the ISOs and used these data to evaluate the threat posed by the six different developmental groupings that terminated in each session. Thus, the DMs estimated the probability (P) of a particular type of deployment (H) given the probabilistic attribute data (D) describing the characteristic of the developmental grouping. Each of the eight DMs performed this task independently even though they all received the same information from the ISOs. Each grouping was classified into one of eight different hypotheses [P(H/D)] which described the possible strategies to be employed by aggressor in initiating some hostile action across the borders of his homeland area. These hypotheses were simply labeled alphabetically A through H.

The DM provided another type of probabilistic estimate, namely, the probability that the various attribute data (D) would be observed if H_i (a particular hypothesis) were true. These P(D/H) estimates were used in the Bayesian solution to obtain the model's P(H/D) estimate.

Design. Eight DMs were assigned to two groups of four each, which were distinguished on the basis of the kind of feedback provided (see Table 1): Group I received knowledge of the accuracy of their own judgments only, and Group II received knowledge of the accuracy of their own judgments and of those provided by a Bayesian process as well. For both groups the judgments of interest were P(H/D) for eight hypothesis categories. The eight hypothesis categories, A through H, were divided into four groups of two each. As illustrated in Table 1, feedback was never given when E or C was correct but was given 33% of the time when Category A or F was correct, 67% of the time when D or G was correct, and 100% of the time when H or B was correct. There was a total of 86 developmental groupings presented for each level of percentage of KR.

RESULTS AND DISCUSSION

The results were analyzed using two types of response measures: one, a verified certainty score (VCS) which is the actual probability placed in the correct hypothesis category; and the other, a dichotomous scoring procedure (DS) where *S* is given full credit if the highest probability is placed in the right hypothesis category and no credit at all if the highest probability is placed in one of the wrong hypothesis categories. If a tie occurred between several hypothesis categories and one of these was correct, proportional credit was given.

The analysis of variance of the VCS and of the DS indicated the same pattern of significant effects. The similarity of the results for the two response measures is corroborated by an examination of Figures 2 and 3 showing similar functions for the two dependent measures.

The effects of the changes in percentage of KR provided the most pertinent information in this study—for VCS, $F(3/1026) = 152.97$, $p < .001$; for DS, $F(3/1026) = 167.29$, $p < .001$. The data were further analyzed by a

Newman-Keuls test which indicated that the main effect (for both types of dependent measures) occurred between 0% and 33% KR ($p < .01$) with no differences between 33%, 67%, and 100% KR. The data support the notion that human beings, even in very complex situations, do not necessarily need 100% KR to sustain performance. Though the necessary amount of feedback is a very task-specific question, there are many situations where it is very difficult to provide anything like 100% KR and it is encouraging to note that there is some evidence developing which points to a critical level somewhere between 0% and 33% KR.

In earlier studies Schum et al. (1966) found that the human P(H/D) responses using dichotomous scores were often equal (for high-fidelity conditions) but never superior to the Bayesian model. For low-fidelity conditions, the model was consistently superior in its first-choice estimates of the correct-hypothesis category. This indicated that more consistency was present in the decision makers' estimates of P(D/H) than they were able to use in estimating P(H/D). The model was able to aggregate the human P(D/H) estimates and consequently was superior on

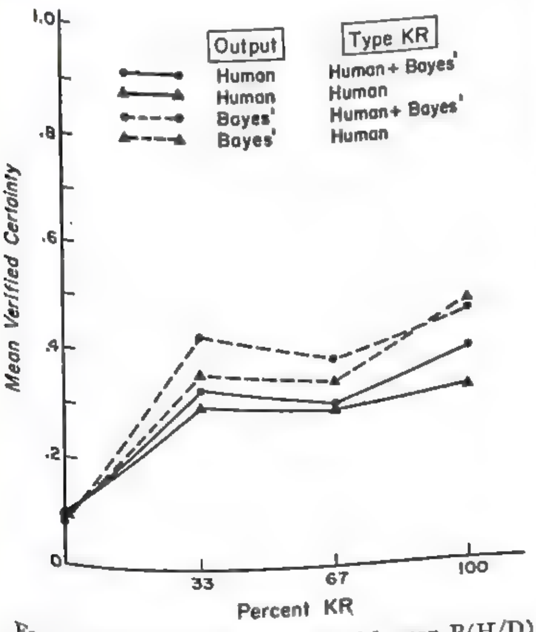


FIG. 2. Human and Bayesian model mean P(H/D) estimates as a function of percentage of KR and specificity of KR.

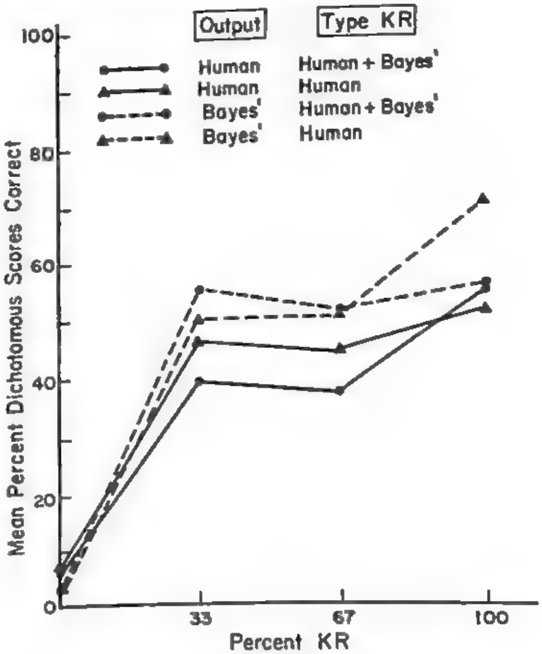


FIG. 3. Human and Bayesian model mean percent dichotomous scores correct as a function of percentage of KR and specificity of KR.

the basis of first-choice estimates. Using verified certainty scores, previous studies (Schum et al., 1966) indicated that human decision makers were able on occasion to provide superior P(H/D) estimates. One of the reasons advanced to explain these data was that *Ss* could, without penalty, maximize P(H/D) in hypothesis categories they considered likely to be correct. However, as occurred with the dichotomous measures, the human verified certainty scores were inferior to the Bayesian model for reduced fidelity conditions. In the present study the dichotomous and verified certainty scores were consistent in the output analysis. That is, the model estimates of P(H/D) were consistently superior to the human data for the 33%, 67%, and 100% KR conditions—for VCS, $F(1/342) = 69.67$, $p < .001$; for DS, $F(1/342) = 23.93$, $p < .001$. If the relationship between high fidelity of information and high percentage of KR could be equated, the humans would have been expected to perform as well as the model for the higher percentage of results conditions. Obviously this did not occur, and it seems likely that the large number of developmental groupings on which the decision maker

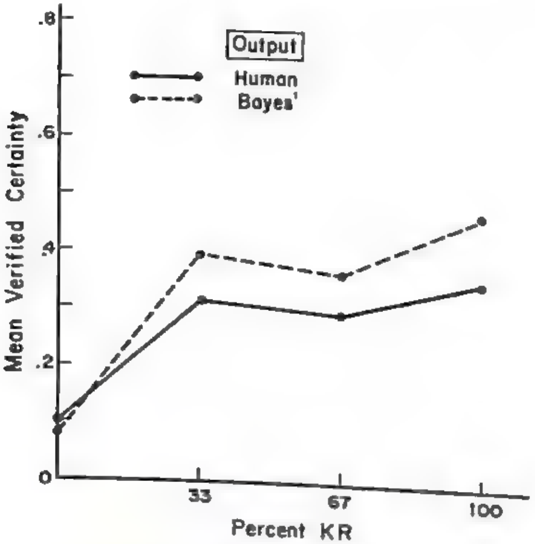


Fig. 4. Human and Bayesian model mean P(H/D) estimates as a function of percentage of KR.

did not receive feedback (the 0%, 33%, and 67% groups) prevented a complete understanding of the type of developmental groupings each P(H/D) hypothesis could possibly include. Thus, the lack of complete feedback resulted in an ambiguous situation which has been shown to result in superior model estimates of P(H/D).

The interaction between percentage of KR and the output—for VCS, $F(3/1026) = 13.67$, $p < .001$; for DS, $F(3/1026) = 6.17$, $p < .001$—again pointed to the importance of the percentage of KR condition between 0% and 33%. An examination of Figure 4 showing the verified certainty scores (the dichotomous relationship was similar) supported this contention. The model and human responses at the 0% KR level overlapped. Without any feedback, neither the human nor the model could provide good estimates from the P(D/H) scores. The triple interaction—for VCS, $F(3/1026) = 4.39$, $p < .01$; for DS, $F(3/1026) = 5.42$, $p < .01$—as seen in Figures 2 and 3 indicated that the interaction between percentage of KR and output changed as a function of the type of feedback.

The effects of specificity of feedback were not significant, indicating that the extra feedback given by the Bayesian model to the DMs did not provide more accurate judgments than feedback from the DM's own response. Yet, as indicated by the significant output effect, the model did produce superior P(H/D) responses. It appears that *S* was not able to use the extra information provided by the model or, as one experimenter (Annett, 1961) has put it, the number of discriminable response categories provided by the KR did not match *S*'s ability to discriminate. There is one other possible explanation. A report of a previous experiment in this series (Briggs & Schum, 1965) exploring the use of the Bayesian model as an aid indicated that decision makers were unwilling to rely on the model solution and, in the final analysis, preferred their own judgments. This would not appear to be the case in the present study where the DMs had much more extensive training and were very familiar with the optimal nature of the Bayesian solution. For this reason, it appears more likely that the DMs were not able to use the extra feedback provided by the model. In either case, further exploration in a wide variety of situations is necessary in order to extend the generality of the effects of specificity of KR and percentage of KR.

REFERENCES

Annett, J. The role of knowledge of results in learning: A survey. *US Naval Training Device Center Tech. Rep.*, May 1961, No. 342-3.

Arps, G. F. A preliminary report on work with knowledge versus work without knowledge of results. *Psychological Review*, 1917, **24**, 449–455.

Briggs, G. E., & Schum, D. A. Automated Bayesian hypothesis-selection in a simulated threat-diagnosis system. In J. Spiegel & D. E. Walker (Eds.), *Second congress on the information system sciences.* Washington: Spartan, 1965. Pp. 169–176.

Crafts, L. W., & Gilbert, R. W. The effects of knowledge of results on maze learning and retention. *Journal of Educational Psychology*, 1935, **26**, 177–187.

Dodson, J. D. Simulation system design for a TEAS simulation research facility. *Planning Research Corporation Rep.*, 1961, No. R-194.

Eriksen, C. W. Effects of practice with or without correction on discrimination learning. *American Journal of Psychology*, 1958, **71**, 350–358.

Gilliland, A. R. The effect of practice with and without knowledge of results in grading handwriting. *Journal of Educational Psychology*, 1925, **16**, 532–536.

Green, R. F., Zimiles, H. L., & Spragg, S. D. S. The effects of varying degrees of knowledge of results on knob-setting performance. *US Naval*

Special Device Center Tech. Rep., 1955, No. 241-6-20.

KIMBLE, G. A. *Hilgard and Marquis' conditioning and learning.* New York: Appleton-Century-Crofts, 1961.

LINDAHL, L. G. Movement analysis as an industrial training method. *Journal of Applied Psychology,* 1945, 29, 420–436.

MICHELS, K. M. The effects of fixed ratio random reinforcement on response latency of monkeys. *Journal of Comparative and Physiological Psychology,* 1955, 48, 32–36.

SCHUM, D. A., GOLDSTEIN, I. L., & SOUTHARD, J. F. The influence of experience and input information fidelity upon posterior probability estimation in a simulated threat-diagnosis system. *USAF AMRL Tech. Rep.,* April 1965, No. 65-25. (a)

SCHUM, D. A., GOLDSTEIN, I. L., & SOUTHARD, J. F. Further investigation of the effects of reduced input data fidelity upon the determination of posterior probabilities in a simulated threat-diagnosis system. *USAF AMRL Tech. Rep.,* December 1965, No. 65-233. (b)

SCHUM, D. A., GOLDSTEIN, I. L., & SOUTHARD, J. F. Research on a simulated Bayesian information-processing system. *IEEE Transactions on Human Factors in Electronics,* 1966, HFE-7, 37–48.

(Received August 25, 1966)

Manuscripts Accepted for Publication in the *Journal of Applied Psychology*

Influence of Favorable and Unfavorable Information upon Assessment Decisions: James W. Miller* and Patricia M. Rowe: Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada.

Underlying Dimensions of Personal Background Data and Their Relationship to Occupational Classification: Melany E. Baehr* and Glenn B. Williams: Industrial Relations Center, University of Chicago, 1225 East Sixtieth Street, Chicago, Illinois 60637.

Interactions among Interests, Abilities, and Career Plans: William W. Cooley*: Director, Project TALENT, 135 North Bellefield Avenue, Pittsburgh, Pennsylvania 15213.

Response Tendencies in the SVIB: The Popular, The Rare, and The Socially Desirable: Donald G. Zytowski* and James A. Walsh: Student Counseling Service, 101 Building H, Iowa State University, Ames, Iowa 50010.

Scaling Assumptions Underlying Weighting in Job-Classification Systems: Frank J. Landy* and Arthur J. Elbert: Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43402.

Differences in Opinion-Survey Response Patterns as a Function of Different Methods of Survey Administration: John R. Hinrichs* and Robert D. Gatewood: Personnel Research, IBM Corporation, 112 East Post Road, White Plains, New York 10601.

Convergent and Discriminant Validation of Satisfaction and Desire Measures by Interviews and Questionnaires. Clayton P. Alderfer*: Graduate School of Business and Public Administration, Cornell University, Ithaca, New York 14850.

Self-Esteem of Industrial Workers: Joel Lefkowitz*: 63 East 9th Street, New York, New York 10003.

Decision Models for Credit Scoring System Cut-off Determination: James H. Myers*: Department of Marketing and Business Communications, University of Southern California, University Park, Los Angeles, California 90007.

(Continued on page 361)

Journal of Applied Psychology
1967, Vol. 51, No. 4, 352–356

INFLUENCE OF INTERPOLATED PERIODS OF ACTIVITY AND INACTIVITY UPON THE VIGILANCE DECREMENT [1]

WILLIAM BEVAN, LLOYD L. AVANT, AND HARRY G. LANKFORD

Kansas State University

4 independent groups were observed in a simple visual detection task. The control group, which monitored the display continuously for 90 min., suffered a reliable decrement in performance during the course of observation. 1 experimental group engaged in vigorous physical exercise for 5 min. after each 30 min. of watch-keeping, the 2nd group solved anagrams for 5-min periods, and the 3rd was subjected to 5-min periods of sensory restriction. All of the experimental groups performed the vigilance task at a high level—with no decrement—throughout. The implications of these results for the significance of successive change in input are discussed.

This simple experiment is a recapitulation of sorts. It examines, under a *common* set of conditions, the effect both of periods of activity and of inactivity upon signal detection, for the light its data may throw upon the activation theory of vigilance.

First suggested by Deese (1955), and elaborated on a growing background of information about the Reticular Activating System, this theory postulates a curvilinear relationship between alertness and the level of central nervous system (CNS) activity produced by overall sensory input and/or central feedback, with maximum alertness being associated with some intermediate level of activity. It accommodates a variety of empirical results: superior performance associated with high as contrasted to low signal rates (e.g., Kappauf & Powe, 1959), with the addition of noncritical signals (e.g., Weiner & Ross, 1962), with greater response complexity (e.g., Adams, Stenson, & Humes, 1961), with dual-mode redundant tasks (e.g., Buckner & McGrath, 1961), with normal versus reduced ambient stimulation (e.g., Myers, Murphy, Smith, & Windle, 1962), and, of course, with more intense and larger stimuli (e.g., Adams, 1956).

But, the literature is not wholly supportive. For example, Alluisi and Hall (1963) view a decrement in auditory vigilance during periods of high multiple-task activity as evidence against the theory. However, the absence of control observations under a single-task regimen, with the possibility of a larger attendant decrement, weakens the significance of their argument. In contrast, Bakan and Manley (1963) report fewer auditory detections in sighted than in blindfolded *S*s. Here one may speculate, along with the authors, either in terms of distractibility in the presence of a more-difficult-than-usual vigilance task or in terms of greater-than-optimum arousal in the presence of a complex task. In an experiment designed specifically to test the activation theory, Adams and Boulter (1962) required their *S*s to make a discrimination response under conditions involving different magnitudes of head and eye movement (proprioceptive feedback) and short-term memory (central feedback). However, when one notes the overall task requirements and the use of paid *S*s, one may question that the experimental procedures effected significant differences in arousal level.

A plausible alternative technique for producing different prevailing levels of alertness involves interpolating periods into the monitoring program that may be assumed to represent levels of input or feedback that are effectively different from that associated with the vigilance task. Studies dealing with the effect of interpolated activity already exist in the literature. However, they have involved tasks of such varied complexity presented

[1] Supported by Contract Nonr 3634(01) between Kansas State University and the Physiological Psychology Branch, Office of Naval Research. William Bevan and Lloyd L. Avant are now at The Johns Hopkins University.

under such a variety of conditions that their significance as operations for inducing different levels of CNS arousal remains obscure. For example, in the pioneering monograph on vigilance, Mackworth (1950) describes interpolating a telephone message at the midpoint of the experimental session. However, the telephone message was an exhortation to do better during subsequent testing. In contrast, McCormack (1958) studied the effect of interpolated rest upon vigilance performance, but rest consisted of reading popular magazines and the opportunity to smoke. Accordingly, it was decided to select a simple, widely used type of vigilance task (visual detection) and a representative signal rate (24 signals/hr) and to examine the effect of interpolated periods of enhanced sensory input (proprioceptive input from vigorous physical exercise), of increased mental activity (making English words from four-letter anagrams), and of stimulus deprivation (*S* reclined, blindfolded and earmuffed) upon vigilance performance. Following the activation hypothesis, it was expected that the first two conditions would produce a reduction in the vigilance decrement and the last an enhancement.

METHOD

A requirement crucial to the success of this experiment involved identifying a task for which a performance decrement might be reliably produced. Preliminary observations were made with a single lamp mounted behind a small aperture in a homogeneous black screen, *S* being requested to signal whenever he saw it illuminated. Different intensities of illumination were tried without success. Next, the task was changed unsuccessfully to require response to the aperiodic absence of a periodically appearing light. A satisfactory task was finally found in an adaptation of a procedure used by Baker (1960) which introduced an element of search into the setting. The *S* was asked to report a flash of light when it appeared at any one of the corners of a square subtending a visual angle of slightly greater than 1°. The design was a simple multiple-group design, *Ss* being assigned at random to either the control or one of the three experimental conditions.

Subjects. Four groups of 10 males each, volunteers from the Introductory Psychology course at Kansas State University, served as *Ss*.

Apparatus and procedure. The *Ss* were tested individually. The *S* and *E* sat on opposite sides of a table, separated from each other's view by a large black wooden screen. On *S*'s side a wooden apron extended from the screen to the edge of the table at an angle of about 30°. The visual display consisted of four 2-mm. holes at the corners of a 1-in. square, backlit by 6-V. signal lamps (General Electric 51) filtered through Wratten A-25 filters. The display was at eye level and approximately 20 in. from *S*. The duration of the signal was ½ sec. The schedule of presentation intervals for each half-hour period of monitoring was a popular one, first used by Mackworth (1950): ¾, ¾, $1\frac{1}{2}$, 2, 2, 1, 5, 1, 1, 2, 3, and 10 min. in that order. The presentation of signals, including their duration, was programmed by means of Hunter decade interval timers. Their spatial location on successive trials was randomly assigned.

Upon entering the dimly lit test room, *S* was asked to surrender his watch and then was seated before the display. He was told that the experiment was a study in signal detection and instructed to move a silent low-friction toggle switch from left to right whenever he saw a red light behind one of the small apertures in the screen. This action illuminated a small signal lamp on *E*'s side of the screen and a response was recorded. After each response, *S* immediately reset the response switch to the off position. The regimens for the several groups were as follows:

(*a*) Control: this group performed the monitoring task continuously for 90 min. and received a total of 36 signals.

(*b*) Experimental-physical exercise: this group monitored the display for three 30-min. periods separated by two 5-min. periods of exercise, during which *S* was required to pedal a Sears Roebuck exercycle at speeds between 10 and 15 mph with a load setting of from 6,000 to 8,000 gm/cm torque.

(*c*) Experimental-mental activity: this group monitored the display for three 30-min. periods separated by two 5-min. periods during which *S* was required to supply one English word for each four-letter anagram presented. Two 25-item lists of anagrams were available for use during each work period.

(*d*) Experimental-stimulus deprivation: this group monitored the display for three 30-min. periods separated by two 5-min. periods during which *S* reclined quietly on a cot, wearing plastic goggles, of the type worn by X-ray technicians but rendered opaque by several coats of flat black paint, and Willson sound barrier ear muffs.

Each *S*'s performance was scored in terms of the number of correct detections per 30-min. period, and statistical tests were performed upon these data. Following convention in the presentation of vigilance data they were transformed to percentage of correct detections for presentation in Figure 1.

RESULTS AND DISCUSSION

The vigilance decrement. Inspection of Figure 1 indicates a decrement in control-group performance across successive 30-min. periods of monitoring. While it is not as pro-

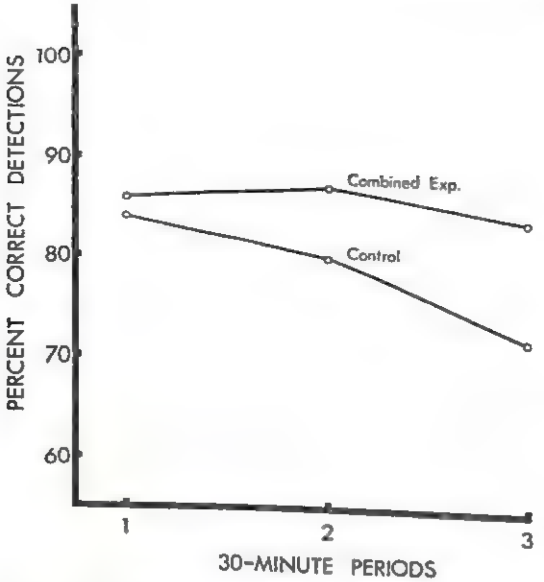


FIG. 1. Vigilance performance on successive half hours of a visual detection task when observation was continuous (control group) or interrupted by 5 min. of physical activity, problem solving, or reduced stimulation (combined experimental groups).

nounced as that observed in some other studies, extending only from an average of 84% for the initial period to 72% at the end of the final period, it is consistent and thoroughly reliable (the F ratio for between periods in the control group $= 9.69$; with df of 2 and 18, $p < .01$). Since the signals all fell well within the region of clear vision, were well above the chromatic threshold, and spaced at intervals minimally of almost a minute's duration, it is reasonable to conclude that the decrement is associated with central—in contrast to peripheral—mechanisms.

Performance in the experimental groups. Examination of Figure 1 also quickly indicates two things about the influence of the several experimental treatments upon the vigilance decrement: they resulted in a facilitation of performance and produced a decrement-free record—F ratio for (Control versus Combined Experimental) $\times$ Periods $=$ 3.73; with df of 2 and 72, $p < .05$—and they did not differ among themselves in their effect upon the decrement—F (Between E's) $\times$ Periods $(2/72) = 1.03$; $p > .05$. Thus the results for the two groups experiencing interpolated activity are in line with the hypothesis stated above while those for the one group subjected to brief periods of reduced stimulation are opposite of expectation. These latter data, which are consistent with a number of previous studies reporting a facilitating effect for rest (e.g., Jenkins, 1958; McCormack, 1958; Saldanha, 1955; Solandt & Partridge, 1946), raise serious question for an explanation of vigilance based on a simple positive relation between alertness and intensity or amount of physical stimulation, irrespective of whether or not level of stimulation is associated with the frequency, intensity, or duration of signals or with response activity generated by the task. Indeed, a decrement was reported by Whittenberg, Ross, and Andrews (1956) who required Ss to respond to all jumps of the Mackworth clock, pushing a toggle switch in one direction for the double-jump signals and in the opposite direction for the single-jump nonsignal events. That level of activity per se does not account for vigilance is further demonstrated by Baker's (1963) observation that as performance declined, general activity or restlessness increased.

A comparison of the three experimental treatments indicates them to have at least one potentially significant characteristic in common: the interpolated period represents a change in an otherwise invariant routine. This suggests variability rather than level to be the significant property of stimuli for vigilance. Consistent with the *physiological* principle of stimulation, Hebb (1955), in the delineation of his conceptual nervous system, assigned great importance to stimulus variation for arousal, and Scott (1957), in an extension of Hebb directly relevant to the problem of the vigilance decrement, identified a reduction in nonspecific stimulus effects (i.e., arousal) with S's continued exposure to input. This he called sensory habituation. Data like those of Kirk and Hecht (1963) are beautifully tailored to fit the stimulus variation theory of vigilance. Although the presentation of a constant noise above the ambient level of the test room had no effect upon the probability of signal detection in a visual monitoring task, the addition of variable noise of the same average intensity reliably enhanced detection performance. This approach provides ready conceptual accommodation for reports

of enhanced performance with multimodal redundant signals (e.g., Buckner & McGrath, 1961), discriminable artificial signals (Wilkinson, 1964), and increased response complexity (e.g., Adams, Stenson, & Humes, 1961). However, one is still faced with results like those of Adams and Boulter, Alluisi and Hall, and Whittenberg et al., cited above, and of Baker and Harabedian (1962), who found that the simultaneous presentation of a second task had no influence upon vigilance performance, and of Luce (1964), who observed that response complexity had no facilitating influence.

A review of these last-mentioned studies yields the impression that in conceptualizing variability one must distinguish between change and variety, that is, successively occurring as distinct from simultaneously occurring variation. The literature on interpolated regimens, these last unequivocally successful as a technique for enhanced performance, is a literature dealing with *change*. A recent experiment by Gruber (1964) underlines this point. Although the simultaneous presentation of visual and auditory signals yielded a generally higher level of detection than either type alone, the introduction of alternation between the two types produced a strikingly better level of performance than under any other regimen including the combined. Recognition of the importance of *successive* stimulus variation directs attention to the *incentive* value of change; change is important because it breaks the monotony of a repetitive task. The observation of the vigilance decrement and the salutary effect of a periodic change in activity is reminiscent of Karsten's (1928) observation of the deterioration of performance in a repetitive task and rapid recovery with a change in set. Relevant also are Butler's (1953, 1958) observations of the powerful reinforcement value that opportunities for viewing or listening to varied stimuli have for animals confined in monotonous visual or auditory environments. Specific to the vigilance setting are Bevan and Turner's (1965, 1966) report of a reduction in errors of omission with anticipation of a change from positive reinforcement for correct detections to negative reinforcement for errors or vice versa. In addition, there are the observations of Smith, Lucaccini, Groth, and Lyman (1966) that detection performance not only was clearly better when *S* monitored the display intermittently with an auditory cue marking the beginning and end of the vigilance period than when monitoring was continuous. Performance showed improvement rather than decrement throughout the test period. Furthermore, when the off periods were occupied by an anagram task, *S*s greatly underestimated the duration of the vigilance task and judged it to be interesting. Relative to these results are the earlier data of McGrath (1960) who reported superior visual detection with a background of sequential periods of music, recorded conversation, and mechanical noise, as compared to a continuous background of white noise and, of course, Mackworth's (1950) enhancement of performance following the interpolation of an exhortatory telephone message.

With the recognition of the incentive properties of stimulus change we have come full circle to Deese's view of arousal as an adjunct to expectancy theory.

REFERENCES

ADAMS, J. A. Vigilance in the detection of low-intensity visual stimuli. *Journal of Experimental Psychology*, 1956, **52**, 204–208.

ADAMS, J. A., & BOULTER, L. R. An evaluation of the activationist hypothesis of human vigilance. *Journal of Experimental Psychology*, 1962, **64**, 495–504.

ADAMS, J. A., STENSON, H. H., & HUMES, J. M. Monitoring of complex visual displays: II. Effects of visual load and response complexity on human vigilance. *Human Factors*, 1961, **3**, 213–221.

ALLUISI, E. A., & HALL, T. J. Declines in auditory vigilance during periods of high multiple-task activity. *Perceptual and Motor Skills*, 1963, **16**, 739–740.

BAKAN, P., & MANLEY, R. Effects of visual deprivation on auditory vigilance. *British Journal of Psychology*, 1963, **54**, 115–119.

BAKER, C. H. Maintaining the level of vigilance by means of artificial signals. *Journal of Applied Psychology*, 1960, **44**, 336–338.

BAKER, C. H. Further toward a theory of vigilance. In D. N. Buckner & J. J. McGrath (Eds.), *Vigilance: A symposium*. New York: McGraw-Hill, 1963. Pp. 127–154.

BAKER, C. H., & HARABEDIAN, A. Performance in an auditory vigilance task while simultaneously tracking a visual target. Los Angeles: Human Factors Research, Inc., Technical Report No. 740–2, 1962.

BEVAN, W., & TURNER, E. D. Vigilance performance with a qualitative shift in reinforcers. *Journal of Experimental Psychology*, 1965, **70**, 83–86.

BEVAN, W., & TURNER, E. D. Vigilance performance with a qualitative shift in verbal reinforcers. *Journal of Experimental Psychology*, 1966, **71**, 467–468.

BUCKNER, D., & MCGRATH, J. J. A comparison of performance on single and dual sensory mode vigilance tasks. Los Angeles: Human Factors Research, Inc., Technical Report No. 8, 1961.

BUTLER, R. A. Discrimination learning by rhesus monkeys to visual-exploration motivation. *Journal of Comparative and Physiological Psychology*, 1953, **46**, 95–98.

BUTLER, R. A. The differential effect of visual and auditory incentives on the performance of monkeys. *American Journal of Psychology*, 1958, **71**, 591–593.

DEESE, J. Some problems in the theory of vigilance. *Psychological Review*, 1955, **62**, 359–368.

GRUBER, A. Sensory alternation and performance in a vigilance task. *Human Factors*, 1964, **6**, 3–12.

HEBB, D. O. Drives and the CNS (conceptual nervous system). *Psychological Review*, 1955, **62**, 243–254.

JENKINS, H. M. The effect of signal-rate on performance in visual monitoring. *American Journal of Psychology*, 1958, **71**, 647–661.

KAPPAUF, W. E., & POWE, W. E. Performance decrement at an audiovisual checking task. *Journal of Experimental Psychology*, 1959, **57**, 49–56.

KARSTEN, A. Psychische Sättigung. *Psychologische Forschung*, 1928, **10**, 142–254.

KIRK, R. E., & HECHT, E. Maintenance of vigilance by programmed noise. *Perceptual and Motor Skills*, 1963, **16**, 553–560.

LUCE, T. S. Vigilance as a function of stimulus variety and response complexity. *Human Factors*, 1964, **6**, 101–110.

MACKWORTH, N. H. Researches on the measurement of human performance. London: Medical Research Council, Special Report Series No. 268, 1950, pp. 22–25.

MCCORMACK, P. D. Performance in a vigilance task as a function of inter-stimulus interval and interpolated rest. *Canadian Journal of Psychology*, 1958, **12**, 242–246.

MCGRATH, J. J. The effect of irrelevant environmental stimulation on vigilance performance. Los Angeles: Human Factors Research, Inc., Technical Report No. 6, 1960.

MYERS, T. I., MURPHY, D. B., SMITH, S., & WINDLE, C. Experimental assessment of a limited sensory and social environment: Summary results of the HumRRO program. Monterey, Calif.: HumRRO Research Memo., February 1962.

SALDANHA, E. L. Alternating an exacting visual task with either rest or similar work. Cambridge: Medical Research Council, APU 289/97, 1955.

SCOTT, T. H. Literature reviewing the intellectual effects of perceptual isolation. Defence Research Board, Department of National Defence, Canada. Report No. HR66, 1957.

SMITH, R. L., LUCACCINI, L. F., GROTH, H., & LYMAN, J. Effects of anticipatory alerting signals and a compatible secondary task on vigilance performance. *Journal of Applied Psychology*, 1966, **50**, 240–246.

SOLANDT, D., & PARTRIDGE, D. M. Research on auditory problems presented by Naval operations. *Journal of Canadian Medical Services*, 1946, **3**, 323–329.

WEINER, H., & ROSS, S. The effects of "unwanted" signals and D-amphetamine sulfate on observer responses. *Journal of Applied Psychology*, 1962, **46**, 135–141.

WHITTENBERG, J. A., ROSS, S., & ANDREWS, T. G. Sustained perceptual efficiency as measured by the Mackworth Clock Test. *Perceptual and Motor Skills*, 1956, **6**, 109–116.

WILKINSON, R. T. Artificial 'signals' as an aid to an inspection task. *Ergonomics*, 1964, **7**, 63–72.

(Received September 16, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 4, 357–361

WORKER PREFERENCES AMONG TIME-OFF BENEFITS AND PAY

STANLEY M. NEALEY AND JAMES G. GOODALE

University of Illinois

197 industrial workers expressed their preferences among 6 proposals for additional paid time off the job. Preference for a comparable pay raise was also measured. Extra vacation was most preferred while a proposal to shorten the workday was least preferred. The pay raise was 5th in preference. Differences in preference were related to sex, age, marital status, and job satisfaction. Foremen were able to predict overall worker preferences with high accuracy.

The level of unemployment in the United States during the past decade has resulted in pressure to spread the work by reducing hours of work per employee. This new emphasis on reduction in work time is a continuation of a trend that stretches back to the emergence of the Industrial Revolution, when 16-hr. days were not uncommon. The steel industry in the United States dropped the 12-hr. day only in 1923. The average nominal workweek has steadily declined, from 55 hr. in 1914 to 50 hr. in 1920 (National Industrial Conference Board, 1925). In 1938, the Fair Labor Standards Act defined the statutory workweek at 40 hr. Continuing this trend, Local 3 of the International Brotherhood of Electrical Workers (AFL-CIO) in New York City secured the 25-hr. week in 1962. While the 25-hr. workweek may be a long time coming into general use, it seems clear that the 40-hr. week is no more sacred in the '60s than was the 48-hr. week in the '30s.

Reductions in work time have taken a variety of forms. Henle (1962) reports that paid vacation increased on the average by 6 days and paid holidays by 4 days between 1940 and 1960. Increased relief time, industrial sabbaticals, and paid early retirement have accounted for additional reductions in work time, all without reduced earnings.

With automation and population both expanding, the trend to reduced hours of work per employee will almost surely continue. The forms it will take are, however, difficult to predict. Reducing the 12-hr. day was a more salient goal in the 1920s than alternatives like extra paid holidays or early retirement. Today, prosposals to reduce the 8-hr. day do not seem to enjoy the same clear advantage when compared with a 3-mo. sabbatical every 5 yr. or an extra week of vacation. In other words, the form that further reductions in work time should take may depend more on employee preferences than humanitarian goals.

Yet there is no evidence in the published literature that employee preferences among the various forms of time off are being measured directly. Industry negotiators often see their role as a reactive one with respect to labor demands, and as such they tend to place the responsibility for knowing what the workers want on the union. It seems clear, however, that most unions do not systematically measure rank-and-file preferences for alternative forms of compensation.

Nealey (1964a) used the paired-comparison method to measure the preferences of industrial workers for six compensation options including more pay, increased vacation, hospital insurance, and more pension. Even though the options were of equal economic cost, they varied markedly in overall preference. In addition, significant variations in preference were related to demographic variables. In subsequent studies, Nealey (1963, 1964b) confirmed that preferences for compensation options varied markedly between demographic subgroups, from plant to plant in a single company, and even with the level of job satisfaction.

The present study used the methodology developed in these previous studies to measure employee preferences among a set of compensation options, all of which involve additional paid time off the job. In addition, foremen predictions of preferences were compared to the actual preferences.

METHOD

The *Ss* were 210 blue-collar workers and 53 foremen from the parts department of a heavy equipment company located in a small town in the Midwest. The workers were randomly selected from among the approximately 700 employees on the first two shifts. The foremen group included virtually all foremen from the first two shifts. Responses from 13 workers were improperly completed and thus discarded. Results will be reported for the remaining 197 workers (180 males and 17 females). All workers were members of the United Auto Workers (AFL-CIO); most were involved with order filling and parts handling. Wages were quite homogeneous across the sample.

A paired-comparison questionnaire measuring preferences for six time-off and one pay option was administered by the authors to groups of 28–30 *Ss* each. Foremen were tested in separate groups. They were asked to complete the questionnaire *as they though the workers would complete it*, thus providing a prediction of worker preference. Questionnaires were administered on company time; *Ss* were asked not to identify themselves. The worker questionnaire also included Kunin's (1955) faces scale, as a measure of overall job satisfaction, and five demographic questions: sex, age, company seniority, marital status, and number of dependent children.

The seven compensation options, as they were presented to *Ss*, are listed below.

A 4-day workweek of 9 hr. and 45 min. each day.
A 5-day workweek with shorter working days of 7 hr. and 50 min.
Five Fridays off per year with full pay. This amounts to five 3-day weekends per year, in addition to any 3-day weekends already scheduled.
Early retirement by accumulating 5 days per year until retirement. Retirement age is 65 minus number of days accumulated, with full pay continuing until age 65.
Leave of 25 working days with full pay every 5 yr. This amounts to 5 wk. off every 5 yr., in addition to normal vacation.
Additional vacation of 1 wk. per year, with full pay, to be added to present vacation period and personal absence pay.
Pay increase of 2%.

Each of the six time-off options involves a 40-hr. reduction in work time per year, with no reduction in yearly salary (40 hr. equals 2% of the work year). The pay option involves a 2% pay raise. Thus, all seven of the options are of equal monetary value.

The options were abbreviated for paired-comparison presentation. The resulting 21 pairs of options were arranged on the questionnaire according to the method developed by Ross (1934) which balances right and left appearance of each option and results in maximal serial separation of options within the list.

In order to further control for time-order error, two forms of the questionnaire were prepared. Pairs of options were arranged on the two forms with right and left sides reversed and in inverted order. One form was completed by 99 workers, the other by 98.

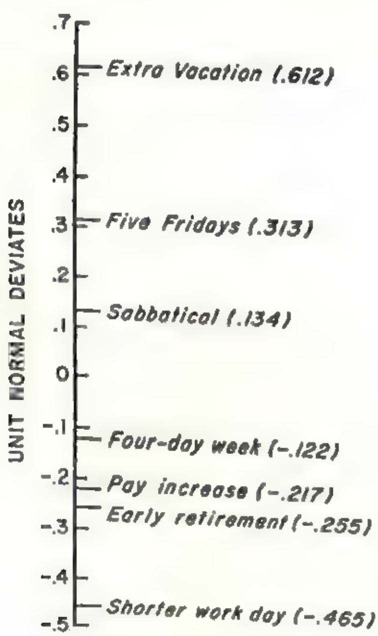


FIG. 1. Relative preference for time-off benefits and pay.

RESULTS AND DISCUSSION

The preferences of the total group of 197 workers were scaled in unit normal deviates and are displayed in Figure 1. The extra week of vacation was most highly preferred while the sorter working day was last in preference. The shorter-work-day option represents time off in its most fractionated form. These data, as well as spontaneous worker comments, indicate that a workday only 10 min. shorter than at present may hardly represent a JND. The 2% pay raise was only fifth in preference, falling behind even the 4-day workweek option with its 9 hr. 45 min. workday. The 4-day week with longer workdays is perhaps the most radical of the options presented. If in even a fraction of situations it proves to be more popular than comparable pay raises, its introduction would mean major changes in industry and perhaps one's whole way of life.

Foremen predictions of preference are compared to the actual worker preferences in Table 1. The data are expressed in terms of

the mean number of times each option was preferred by each S. One-way analyses of variance were performed on each of the options.[1] As Table 1 indicates, only in the case of the pay raise did the foremen predictions differ significantly from worker preferences. The rank-order correlation between the two scales is .90. The accuracy of the present predictions is remarkable when compared to several similar attempts. Raube (1947) had employees rate the importance of 71 morale factors. Considering only the top 10 factors, the rank-order predictions of 50 top executives correlated .30, and those of 42 union officers correlated .20, with the workers' ratings of importance. The National Retail Dry Goods Association (1939) reported that the ranking of importance of eight morale factors by 3,000 retail-stores employees correlated −.21 with the predictions of importance made by several hundred employers. The inability of management to predict employee wants has become part of the lore of industrial psychology. The accuracy of the foremen in the present study would seem to warrant a reevaluation of this verdict.

In the present case, both workers and foremen responded to the same questionnaire, and

[1] Since there is some doubt whether these data meet all the assumptions underlying the use of analysis of variance, a sample of results was analyzed using both analysis of variance and the Mann-Whitney U test. The significance values resulting from the use of these alternative techniques were virtually identical. Analysis of variance was chosen throughout because it avoids the multiple-comparisons problem of setting alpha levels for the Mann-Whitney U test.

TABLE 1

MEAN WORKER PREFERENCES AND FOREMEN PREDICTIONS OF WORKER PREFERENCES

Option	Workers (N = 197)	Foremen predictions (N = 53)	F^a
Extra vacation	4.53	4.24	1.82
Five Fridays	3.82	3.89	.10
Sabbatical	3.36	3.21	.44
4-day week	2.67	2.98	.85
Pay increase	2.45	3.26	7.64*
Early retirement	2.34	2.06	1.25
Shorter workday	1.82	1.36	3.54

^{a} $df = 1/195$.
* $p < .01$.

the options were specific and quantitative rather than general and qualitative. Whether union leaders and higher management, given the present methodology, can predict worker preferences as well as foremen remains a question for further research.

Differences in preference associated with the demographic variables of sex, age, and marital status are displayed in Table 2. In general, preference for the more highly preferred options did not differ very much from subgroup to subgroup. An exception to this was the tendency for the Five-Fridays option, which was second overall in preference, to be less preferred among older workers ($p < .01$). Males showed significantly higher preference for the pay raise than did females, while preference for the shorter workday was higher among females than males. Both of these sex differences seem readily interpretable. Working women have traditionally been

TABLE 2

MEAN WORKER PREFERENCES BY DEMOGRAPHIC SUBGROUP

Option	Sex		F^a	Age in years			F^b	Marital status		F^a
	Males (N = 180)	Females (N = 17)		18–33 (N = 50)	34–49 (N = 83)	50–65 (N = 64)		Single (N = 26)	Married (N = 171)	
Extra vacation	4.53	4.59	.03	4.58	4.30	4.80	2.19	4.42	4.55	.20
Five Fridays	3.84	3.53	.78	4.34	3.72	3.53	5.90**	4.00	3.79	.50
Sabbatical	3.36	3.35	.00	3.08	3.39	3.55	1.54	3.38	3.36	.01
4-day week	2.65	2.88	.18	3.28	2.71	2.14	4.46*	2.50	2.70	.18
Pay increase	2.54	1.53	4.59*	2.16	2.53	2.58	.94	2.50	2.44	.02
Early retirement	2.34	2.47	.10	1.76	2.47	2.64	4.80**	1.69	2.44	4.48*
Shorter workday	1.71	2.65	4.80*	1.80	1.88	1.77	.08	2.50	1.72	5.23*

^{a} $df = 1/195$.
^{b} $df = 2/194$.
* $p < .05$.
** $p < .01$.

TABLE 3

MEAN WORKER PREFERENCES BY JOB SATISFACTION

Option	Job-satisfaction level				F^a
	(Low) 1 and 2 ($N = 23$)	3 ($N = 83$)	4 ($N = 68$)	5 (High) ($N = 23$)	
Extra vacation	4.52	4.39	4.60	4.87	.81
Five Fridays	3.96	3.86	3.97	3.09	3.25*
Sabbatical	3.30	3.40	3.31	3.43	.06
4-day week	2.83	2.99	2.50	1.87	1.84
Pay increase	2.39	2.48	2.22	3.09	1.43
Early retirement	1.83	2.42	2.49	2.17	1.08
Shorter workday	2.17	1.47	1.91	2.48	2.46

^{a} $df = 3/193$.
* $p < .05$.

less aggressive than men about wages, and they may also feel that a job keeps them from home duties too long. Preference for early retirement, as expected, increased significantly ($p < .01$) with age. This option was also more popular with married than with single workers.

In addition to the demographic variables included in Table 2, preferences by seniority level, number of dependent children, and first versus second shift were investigated. Significant changes in preference by seniority paralleled those by age so closely, even to the significance levels involved, that tabling them would serve no purpose. The similarity of results by age and seniority is not surprising since these variables were correlated .68 in the sample. The intercorrelations among all the other demographic variables were not significant. Preferences of first-shift workers did not differ significantly from those of the second shift, nor did preferences vary significantly according to number of dependent children.

The preferences of four subgroups differing in level of job satisfaction are presented in Table 3. The option of Five Fridays was significantly less popular among the most satisfied workers. Differences in preference for the shorter workday were marginally significant.

If preference judgments of the present type are to be useful, they must satisfy the criterion of an ordinal scale, that is, they must display acceptable transitivity (Torgerson, 1958, pp. 27–29). Kendall (1962, Ch. 11) provides a formula for calculating the number of circular triads in a set of paired-comparison judgments. Circular triads serve as a negative index of transitivity. Kendall also provides a probability distribution of circular triads, assuming that preferences are purely random. In 4,137 preference judgments the *Ss* produced a total of only 276 circular triads, a mean of 1.40 per *S*. This value corresponds to a probability of less than .01 that these preference judgments are random. The present judgments may be assumed to possess ordinal scale properties.

Order-error response bias was evaluated by means of the χ^2 test. The left-hand options were chosen in 50.54%, or 2,091, of the 4,137 presentations ($\chi^2 = 0.49$, $p \simeq .50$). Order error was not significant in these judgments.

The results as a whole reinforce the conclusion suggested by earlier research, that no one wage and benefit package will be highly acceptable to more than a subset of employees. This suggests the following two changes in industrial practice: (*a*) Smaller rather than larger bargaining units offer the kind of flexibility in compensation negotiations that could be responsive to local preferences. (*b*) Some degree of individual choice in wage and benefit selection might be beneficial. Various voluntary enrollment insurance and medical care programs offer this feature at present, but they usually involve employee contributions. Offering comparable benefits on a "cafeteria" basis would allow employees to satisfy their preferences much more fully than

is possible at present. Administrative requirements will, of course, serve as constraints on the amount of individual choice possible in any situation. For instance, in the present case, the options of shorter workday, the 4-day week, and five Fridays off, would probably have to be initiated on a plant-wide basis. Some choice between pay, vacation, the sabbatical, and early retirement might be possible on a self-selection basis.

REFERENCES

HENLE, P. Recent growth of paid leisure for U. S. workers. *Monthly Labor Review*, 1962, **85**, 249–257.

KENDALL, M. G. *Rank correlation methods.* (2nd ed.) New York: Hafner, 1962.

KUNIN, T. The construction of a new type of attitude measure. *Personnel Psychology*, 1955, **8**, 65–77.

NEALEY, S. M. Pay and benefit preference. *Industrial Relations*, 1963, **3**(1), 17–28.

NATIONAL INDUSTRIAL CONFERENCE BOARD. *Wages and hours in American industry.* New York: Author, 1925.

NATIONAL RETAIL DRY GOODS ASSOCIATION, Employee Relations Bureau. Your employee wants you to know him. *Personnel Journal*, 1939, **17**, 357–363.

NEALEY, S. M. Determining worker preferences among employee benefit programs. *Journal of Applied Psychology*, 1964, **48**, 7–12. (a)

NEALEY, S. M. Measurement of worker preferences and the determinants of preferences for industrial compensation and benefit programs. Unpublished doctoral dissertation, University of California, Berkeley, 1964. (b)

RAUBE, S. A. Factors affecting employee morale. (Studies in Personnel Policy No. 85) New York: National Industrial Conference Board, 1947.

ROSS, R. T. Optimum orders for the presentation of pairs in the method of paired comparison. *Journal of Educational Psychology*, 1934, **25**, 375–382.

TORGERSON, W. S. *Theory and methods of scaling.* New York: Wiley, 1958.

(Received September 16, 1966)

(Continued from page 351)

Comparative Managerial Role Perceptions in Military and Business Hierarchies: Vance F. Mitchell* and Lyman W. Porter: University of California, School of Business Administration, Berkeley, California 94720.

Vocational Interests of Community Recreation Administrators Using the SVIB: Keith B. Roys, Sr.*: Department of Recreation and Park Administration, Room 215, 909 University Avenue, University of Missouri, Columbia, Missouri 65201.

Copying Alpha and Numeric Codes by Hand: An Experimental Study: R. Conrad* and A. J. Hull: Medical Research Council, Applied Psychology Research Unit, 15 Chaucer Road, Cambridge, England.

Rate of Comprehension of an Existing Teleprinter Output and of Possible Alternatives: E. C. Poulton* and C. Helen Brown: Assistant Director, Medical Research Council, Applied Psychology Research Unit, 15 Chaucer Road, Cambridge, England.

Organizational Correlates of Executive Roles: Raymond A. Katzell*, Richard S. Barrett, Donald H. Vann, and John M. Hogan: Department of Psychology, New York University, 21 Washington Place, New York, New York 10003.

Use of High School Data to Predict Work Performance: Marshall H. Brenner*: Testing and Research Administrator, Dept. 90-16, Unit 36, Plant A-1, Lockheed-California Company, Burbank, California 91503.

A Biographical Inventory for Identifying Creativity in Adolescent Boys: Charles E. Schaefer and Anne Anastasi*: Fordham University, Department of Psychology, Bronx, New York 10458.

Validity of the Guilty Knowledge Technique: The Effects of Motivation: P. O. Davidson*: Department of Psychology, University of Calgary, Calgary, Alberta, Canada.

Social Immaturity, Intellectual Ability, and Adjustive Behavior in College: David Kipnis*: Department of Psychology, Temple University, Philadelphia, Pennsylvania 19122.

Effects of Signal and Stimulus Rates on Vigilance Performance: Harvey A. Taub* and Francis H. Osborne: Veterans Administration Hospital, Irving Avenue and University Place, Syracuse, New York 13210.

* Asterisk indicates author for whom the address is supplied.

Journal of Applied Psychology
1967, Vol. 51, No. 4, 362–368

ACCURACY OF MANUAL ENTRIES IN DATA-COLLECTION DEVICES

WILLIAM A. SMITH, JR.[1]

Computing Laboratory, Lehigh University

This experiment examined the accuracy of manually recorded messages similar to those encountered in field studies on the accuracy of data collection in production information systems. The high efficiency in correcting errors detected at the time a message was recorded and the relative contributions of format and content mistakes to residual error found in field studies were sufficiently universal that they were reproduced and confirmed under laboratory conditions. Under controlled conditions, manual messages which were 3-, 6-, and 10-digits long contributed significantly different quantities of both total and residual errors. Imposition of a time restraint did not affect the total quantity of mistakes, but it did contribute to significant differences in residual errors, affecting both the ability to detect and correct mistakes at point of entry. About ½ of all the mistakes in observed manual messages under field and laboratory conditions were caused by single-digit substitution. Omission of a digit accounted for another 20%. Transposition mistakes were more frequently encountered in the laboratory experiment than in field studies but they were a less important contributor to inaccurate data recording.

The advantages sought in recording production data at the time and place an event occurs are improved accuracy of the data, improved system status and response to change in production activity, and sufficiency of data.

Immediate acceptance and use of data in an information system necessitates substitution of comprehensive system checks for conventional data-processing verification and supervisory review. The system accepting data must anticipate the kinds and volume of errors in the data in order to recognize and correct them. Some errors may require buffering data in a suspended file during analysis phases in order to adjust to variations in volume or quality discovered (Minor & Revesman, 1962; Rook, 1962; Scales & Chapanis, 1954; Swain, 1963). Several field studies were conducted to investigate the accuracy of production data recorded in machine-readable form in order to enhance the system-design and implementation phases.

In so doing, it was necessary to define and classify the kinds of errors associated with data collection in the shop environment. The data collected in three different production situations, one assembly shop and two job shops, were analyzed and the characteristics and quantity of errors were determined. They fell into three major categories:

1. Message format: Items with wrong format that can be detected and screened from system input (wrong message length, illegal characters, equipment malfunction).

2. Message content: Items that have correct form, but can be detected as logically inconsistent (shop status contradictions, unusual quantities, corrections to good entries, wrong machine or operator designation).

3. Event description: Items that have correct form and are logically processable, but prove inconsistent after subsequent entries (omitted entries, extra entries, wrong sequence, late or early entries, failure to correct detected mistakes).

Messages containing mistakes can be identified either at the time entry is attempted in an input terminal or during processing of the data to update records. As an audit of system accuracy, it is also possible to compare the results of routine processing with records kept by parallel methods, with inventory status, or with predetermined test entries. Messages that remain inaccurate, containing

[1] This paper is adapted from a thesis accepted by the faculty of the Graduate Division of the School of Engineering and Science of New York University in partial fulfillment of the requirements for the degree of Doctor of Engineering Science.

mistakes which are undetected or uncorrected, after normal procedures of checking entries and utilizing the data are called "residual errors." They can be recognized only by the special postprocessing analysis, or audit, mentioned previously.

FIELD STUDIES

Data-collection entries were studied intensively for three manufacturing shop situations which were representative of production environments and which afforded ease in evaluating recorded entries. In summary, these sites had the following characteristics:

1. Assembly-line workers manually dialed 2-digit messages into modified telephone sets at their individual work stations. Inspection and test messages, varying 2–12 digits in length, were recorded into similar sets at individual stations. Time and station identification were automatically entered.

2. Machine-layout operators recorded entries in a modified telephone-set terminal for a group of machines (and workers) in a piece-part job shop. Inspectors made entries in similar devices at their work station. The variable length and format entries required a mixture of manually dialed digits and preassigned job and machine identification cards. Time and station identification were automatically entered.

3. In this precision machine job shop, production and inspection entries were made primarily by clerks at a central production-control station. A standard collection terminal marketed by a computer company was used to enter 11 manual digits; preassigned cards containing part, shop order, and employee identification; and time and date automatically.

A composite representation of the distribution of errors enumerated by category is contained in Table 1. These identified mistakes are further characterized in Table 2 by the means of detection. "Entry" refers to recognition of a mistake while the worker is at the input terminal. "Process" refers to reasonableness or status checks during data processing. "Undetected" mistakes are those identified only in postprocessing audits after the data have been used.

TABLE 1

FIELD-STUDY ERRORS BY CATEGORY

	Error category		
	Format	Content	Event description
Distribution of errors	50%	35%	15%
Proportion detected and corrected	90%	65%	20%
Residual error rate of total entries	0.7%	1.9%	1.8%

The majority of mistakes were discovered in the field studies at the time of entry into collection devices and almost all of them were corrected as indicated in Table 2. These were most commonly format mistakes and obvious content discrepancies which are usually reported in describing error rates for data-recording tasks. The lesser number of inaccurate messages which escaped detection at entry resulted in the principal part of uncorrected error remaining in the data. Table 1 indicates that logical and event-description errors were heavy contributors to the residual, or uncorrected and undetected, errors. Failure to record events, or omitted entries, proved to be a special concern in both the accuracy of the system and the design of processing techniques. Such inaccuracies are difficult to measure in field studies and awkward to emulate in laboratory tasks. The residual error rate for messages containing mistakes after detection and correction procedures was consistent at slightly more than 4% of the total number of

TABLE 2

FIELD-STUDY ERRORS BY SOURCE OF DETECTION

	Source of detection		Undetected
	Entry	Process	
Distribution of errors	64%	21%	15%
Proportion detected and corrected	97%	38%	—
Residual error rate of total entries	0.3%	1.9%	2.2%

entries for production and inspection workers in the situations intensively studied.

Individual workers displayed substantial fluctuation in the accuracy of recording in different time periods. This was similar to the within-worker variation reported by Klemmer and Lockhead (1962). Thus, it was not surprising that significant differences were found in frequency of total errors in recorded data between different shifts in the same production situation and between different sites which used contrasted production-control methods. Comparison of residual errors, those remaining after normal detection and correction procedures, did not exhibit the same statistical differences. They were not susceptible to influence by dissimilarities in supervision, message construction, production control, or individual performance. Only obvious discrepancies in recording procedures produced statistically or practically significant differences in residual errors between shifts or production sites for the steady-state situations studied. No clear preference was demonstrated for assignment of activity reporting duties among clerks, group or crew representatives, and individual workers.

METHOD

The Experiment

The field studies were conducted in actual operating environments over which only a minimum of control could be exercised. It was possible that the kinds and frequencies of errors encountered were affected by unusual parameters or variations which were not easily recognized in shop conditions. Further, it was difficult to establish samples at different levels of factors which had the necessary uniform quantity and randomness for analysis.

An experiment was conducted under laboratory conditions to overcome these limitations with respect to some parameters in order to:

1. Determine the accuracy of recorded messages and categorize errors.
2. Analyze the effect of message length and recording time upon accuracy of data.
3. Characterize the kinds of manual digit mistakes made.

Task. Subjects (*Ss*) recorded visually presented numeric messages on a card-punch keyboard. If a mistake was detected, *S* was instructed to void the message with an error indication and to retransmit the entire message.

Stimulus and apparatus. Sets of 60 randomly chosen data words for each message length were visually displayed as black, handwritten, 2-in.-high numbers on a 4-in.-high white card. Each number was singly presented simultaneously to a group of *Ss*. It was left in view while *Ss* recorded and they were instructed to ignore any number not in sight at a particular time. Two sets of numbers existed for each message length. The messages were always presented in the same order, but *Ss* saw each set only once in the trials. The same practice sets were used for all *Ss* in all practice periods. Digits were not grouped in presentation of any message-length displays. All numbers were recorded on the numeric section of a card-punch keyboard. Registration to successive cards was performed semiautomatically under card control except for initial positioning for a trial at different message length. The *Ss* were seated at normal height for a card punch and each had a clear line of sight to the stimulus. Noise and distraction were minimal.

Training and performance of subjects. Thirty right-handed industrial engineering undergraduates were tested in six groups of five each. All had previously used card punches but had no formal or extensive training and no particular skill in card-punch operation. At the beginning of each trial session, instructions were covered on recording and correcting entries. A set of 20 two-digit numbers was used to illustrate these procedures. Then a separate set of 20 two-digit numbers was recorded under timed conditions for practice and a period for questions allowed before every trial period. Practice and learning effects were minimized by these practice runs and by randomization of factors.

Variables. The number of errors for each set of 60 recorded messages was the dependent variable. The independent variables were:

1. Message length, 3 levels—3, 6, 10 digits. These three levels of message length were representative of the range of consecutive manually entered digits commonly used by system designers and allowed by most commercial data-collection devices.

2. Time for entry, 2 levels—limited and unrestricted time. Unlimited time to enter and correct data and also time pressure generated by allowing only $1\frac{1}{4}$ sec/digit for recording were studied.

3. Groups, 6 levels—6 different sets of five individuals. The *Ss* were assigned to the same groups for all trials, thus nesting the individuals within groups.

Experimental design and procedures. Each *S* was tested in six trial runs conducted at two sessions over a 2-wk. period. The sessions for the groups were randomized as to period of the day and the time between trials. Within each trial session, a set of 60 display messages was run for all message lengths in randomly assigned sequence, which assured that every permutation was accomplished for both time periods. The display sets were assigned so that each appeared an equal number of times in both periods of the day and that each group was exposed to both sets. The other factor, time pressure, was assigned in a random manner to minimize the possible repetitive sequences in presentation

TABLE 3

EXPERIMENTAL DESIGN

Trial date	November					
	1	3	4	8	10	11
First period						
Group	A	B	C	E	F	D
Display set	I	II	I	II	II	I
Factor order of runs	T-3 T-6 T-10	T-3 T-10 T-6	U-6 T-3 T-10	U-10 T-3 U-6	U-10 T-6 T-6	T-6 T-10 U-3
Second period						
Group	B	F	D	A	E	C
Display set	I	I	II	II	I	II
Factor order of runs	U-10 U-3 U-6	U-3 U-6 T-10	T-3 U-10 U-6	U-6 U-10 U-3	T-6 U-3 T-10	U-10 T-6 U-3

Note.—T = timed run; U = untimed run. 3, 6, 10 = number-of-characters message.

while assuring that all groups had both a timed run and an untimed run for each message length. Individuals, who were initially arbitrarily selected for groups, were assigned to different card punches and different positions relative to the display on their successive trials. The design accomplished is shown in Table 3. The procedure for a trial session was to run a practice group of 20 two-digit messages to illustrate instructions and correction procedures without any time restriction. Next, a set of 20 two-digit messages was copied on a timed basis. The time restriction was imposed by:

1. Announcing the run to be timed and reminding that an attempt to copy a message should be made only when the message was in view.
2. Experimenter (*E*) timing exposure of each number card by stop watch, allowing $1\frac{1}{3}$ sec/digit.
3. The *E* standing with hand close to card and moving cards with alacrity. The watch was kept in clear view of all *Ss*.

In contrast, on untimed runs, *Ss* were told they had unlimited time and were asked to indicate if they needed more time before *E* continued to display the next card. The *E* allowed a pause after the last *S* finished recording each number.

The trial runs were then conducted in the designed sequences. Brief pauses in the recording process were planned at the beginning of each number set and at regular display intervals within them to allow registration of punch cards.

Null hypothesis. No differences exist in the number of errors observed in recording visually displayed messages, that is, $e_{A,3,U} = e_{B,3,U} \ldots = e_{F,10,T}$, where A = Group A, 3 = 3-digit message, U = untimed run; B = Group B; F = Group F, 10 = 10-digit message, T = time run.

Alternate hypothesis. The quantity of errors in a set of displayed messages is dependent on one or more of the fixed variables identified as message length, time pressure, or group performance.

Statistical test. Analysis of variance was chosen to analyze this nested factorial experiment. Message length was a fixed variable with three levels; time pressure was a fixed variable with two levels; group performance was a fixed variable with six levels; individual performance was a random variable nested within the groups. If message-length or group-performance differences proved statistically significant, Scheffé's (1933) multiple-comparison test for contrasts would show that specific levels contributed to the differences. The total degrees of freedom, after removing 1 for the grand mean, were 179 (5 Individuals per Group $\times$ 6 Groups $\times$ 3 Lengths $\times$ 2 Times).

Significance level. $\alpha = .01$, to ensure practical significance to any differences.

Samples. The grouping and number of individuals were representative of the quantity of associated persons under shop conditions. Each *S* was asked to record 360 messages which represented activity of a production worker over a period ranging from a week to a month. The total of 10,800 displayed messages was substantial for characterizing errors. The 179 degrees of freedom provided adequate sensitivity for resolution of differences despite the confounding of mean square error with expected mean squares for individuals. The number of observed errors was chosen as the measure because of the uniform quantity of displayed messages in each cell.

RESULTS AND DISCUSSION

Quantity of Errors

The laboratory experiment yielded the results shown in Table 4. All enumerated errors were observed in 30 sets of 60 desired entries

TABLE 4

SUMMARY OF MANUAL EXPERIMENTAL ERRORS

	Message length			Overall
	3	6	10	
Total errors				409
Timed	26	69	122	217
Untimed	28	58	106	192
Residual errors				226
Timed	12	48	82	142
Untimed	13	27	44	84
Total entries	3630	3650	3702	

for each cell shown. Although neither total nor residual errors responded to differences in time pressure on the individuals recording 3-digit messages, there was practical significance to the quantity of errors encountered at different message lengths. The 3-digit messages were apparently impervious to time restraint and displayed substantially lower error rates.

The analysis of variance, shown in Table 5, indicates that message length contributed very significantly to the differences in total errors observed. The effects of time pressure did not yield a statistically significant difference in total errors committed, although 6- and 10-digit messages had noticeably fewer errors in untimed runs. Differences in performances by randomly formed groups were not significant, but this was again attributed to variation in performance of individuals nested within groups as indicated by the relatively large associated mean square. The message lengths proved to be significantly different each from the other by Scheffé tests for multiple comparison. Thus, 6-digit messages had significantly more errors than 3-digit messages at the 1% level. Likewise, 10-digit messages had more errors than did 6-digit ones as indicated below.

Residual errors after normal opportunity for detection and correction were subjected to a separate analysis of variance yielding the results in Table 6. Both message-length and time-pressure effects bore significance for residual errors. Group differences again made a reasonably large contribution to the mean square, but this factor was countered by a relatively high mean square for individuals within groups. Comparison of the message-length contrasts by the Scheffé method showed that 3-, 6-, and 10-digit messages did contribute significantly different residual-error quantities. The difference in residual errors for 3-, 6-, and 10-digit manual messages bore significance at the 1% level. When this variable was isolated from the production environment, it proved to contribute to error experience. The choice of message length for manual digits, even in the short span of sizes studied, proved to be a substantial source of variation in both total and residual errors.

One individual S was an outlier in his performance, having quantities of total and residual errors that were several times the amount of the nearest competitor. No specific assignable cause was determined for the un-

TABLE 5

ANALYSIS OF VARIANCE: TOTAL EXPERIMENT ERRORS

Effect	df	MS	F	
Message length (M)	2	127.239	M/MI	= 36.2*
Time (T)	1	3.472	T/TI	< 1
Groups of individuals (G)	5	16.459	G/I	< 1
Individuals within groups (I)	24	22.453		—
M × T	2	1.439	MT/MTI	< 1
T × G	5	7.339	TG/TI	= 1.57
M × G	10	1.939	MG/MI	< 1
M × I	48	3.523		—
T × I	24	4.681		—
M × T × G	10	3.686	MTG/MTI	< 1
M × T × I	48	4.526		—
	179			

Note.—$F_{.001}(2/48) = 8.01$. $F_{.05}(5/24) = 2.62$.
* $p < .001$.

TABLE 6

ANALYSIS OF VARIANCE: RESIDUAL EXPERIMENT ERRORS

Effect	df	MS	F	
Message length (M)	2	42.506	M/MI	= 19.1**
Time (T)	1	18.689	T/TI	= 6.42*
Groups of individuals (G)	5	14.169	G/I	= 1.01
Individuals within groups (I)	24	14.058		—
M × T	2	6.372	MT/MTI	= 2.57
T × G	5	4.329	TG/TI	= 1.49
M × G	10	1.859	MG/MI	< 1
M × I	48	2.237		—
T × I	24	2.903		—
M × T × G	10	1.952	MTG/MTI	< 1
M × T × I	48	2.473		—
	179			

Note.—$F_{.001}(2/48) = 8.01$. $F_{.025}(1/24) = 5.72$. $F_{.05}(2/48) = 3.18$.
* $p < .05$.
** $p < .001$.

TABLE 7

SUMMARY OF EXPERIMENT RESIDUAL ERRORS

	Untimed	Timed	Total
Entries	5508	5474	10982
Errors	192	217	409
Entry detected	109	86	195
Uncorrected	1	11	12
Undetected	83	131	214
Residual	84	142	226
Error rate (Percentage of entries)	1.5%	2.6%	2.1%

usual error rate, but the *S* was a poor and slow reader. His recorded values illustrated further the variability among *S*s and contributed to the individual's within-group mean squares. Although this individual's error rate was high, it was proportionately so in each category.

The imposed time limit did not seem to cause any difference in the time to record a good entry. Both timed and untimed entries took 1 sec/digit or less to record when no errors were detected or corrections attempted. It was obvious during the experiment that the time pressure caused difficulty in correcting detected mistakes. When no restraint was applied and *S* was allowed to pace corrections himself, he frequently took double the $1\frac{1}{3}$ sec/digit yardstick in order to make a correction and ranged as high as 5 sec/digit.

TABLE 8

CONTRIBUTION OF KINDS OF ERRORS TO RESIDUAL

Experiment	Entries	Residual errors	Error rate
	10982	226	2.1%
Format		96	0.9%
Content		127	1.2%
Event		3	Negligible

Therefore, it was not surprising that residual errors were more numerous when time pressure was imposed. However, the time pressure also allowed many more errors to go undetected by *S*s. It appeared to *E* that *S*s worked with greater assurance under timed conditions, but this confidence was proved insupportable by the higher residual-error rate which resulted.

In recording 10-digit numbers, 29 of the 30 *S*s reported that they developed a pattern of transferring digits from the display to the keyboard. Some varied the pattern used, but 10 *S*s recorded in two groups of 5 digits, six recorded in three groups of 4-4-2 digits, and seven utilized some combination of digits grouped 3-3-4.

Sources of Detection

The controlled experiment confirmed that errors detected by the operator during the

TABLE 9

KINDS OF MANUAL MISTAKES: LABORATORY EXPERIMENT

	Message length			Total	Mistake percent
	3	6	10		
Entries	3630	3650	3702	10982	
Mistakes	50	122	223	395	100.0%
Single digit					
Substitution	35	69	99	203	51.4%
Omission	7	19	47	73	18.5
Insertion	4	3	16	23	5.8
Transposition	2	13	14	29	7.3
Possible transposition	1	7	18	26	6.6
Double digits					
Substitution	—	4	7	11	2.8
Omission	—	3	1	4	1.0
Insertion	—	—	2	2	0.5
Miscellaneous (including multiple-digit mistakes)	1	4	19	24	6.1

recording process were effectively corrected. Table 7 shows that only one detected mistake was missed in the untimed entries. The errors associated with the experimental recording were limited to format and content, except for an extra message and two omitted ones under the timed conditions. Thus, event-description mistakes, which contributed substantially to residual error in site studies, were not included in the error rate calculated. The residual errors were distributed as indicated in Table 8. The proportion of entries that contributed to residual errors caused by format and content mistakes is similar to that found in the field-study and in other experiment data (Klemmer & Lockhead, 1962; Minor & Revesman, 1962; Scales & Chapanis, 1954). The difference compared to the field studies lay in the lack of exposure to event-description mistakes in the experiment. It was presupposed that the number of content mistakes would be fewer in the experiment because misidentification and miscounts were not involved. However, no appreciable practical difference was apparent.

Nature of Manual Mistakes

The mistakes observed in the laboratory experiment were analyzed to determine the kind of manipulation recording faults which occurred. Table 9 shows a tabulation of these results which can be instrumental in designing more effective diagnostic checks of data. The total errors enumerated here were only 395, whereas 409 were reported in previous tabulations. Eleven messages were voided when no mistake existed, one duplicate message was recorded, and two entire messages were omitted. Because no digit mistake could be identified in these errors, they were removed from the total considered for the present purpose.

The possible transposition category resulted from a practical problem of interpreting digit mistakes. Assume the following message was recorded:

> 72 + 732 (where "+" is a character to indicate error detection and voids the preceding digits).

The mistake may be interpreted in one of several principal ways:

1. Substitution of a "2" for "3,"
2. Omission of a "3," or
3. Transposition of the "2" and "3" detected before entry of the last digit.

All such mistakes were classified as "possible transposition" to be consistent. This favored the presupposition that transposition of digits was the most common offense in recording data. Nevertheless, substitution and omission of a single digit were the sources of slightly more than two-thirds of the mistakes encountered and were far more dominant than transposition and possible transposition mistakes. These results confirmed the same relative proportion of kinds of digit mistakes identified in the completely manual messages in the field studies and have been reported by Carlson (1963). In the latter cases, transposition mistakes were even less prevalent than in the experiment.

REFERENCES

CARLSON, C. Predicting clerical error. *Datamation*, 1963, 9(2), 34–36.

KLEMMER, H. T., & LOCKHEAD, G. R. Production and errors in two keying tasks: A field study. *Journal of Applied Psychology*, 1962, 46, 401–408.

MINOR, P. J., & REVESMAN, S. L. Evaluation of input devices for a data setting task. *Journal of Applied Psychology*, 1962, 46, 332–336.

ROOK, L. W., JR. *Reduction of human error in industrial production*. Technical Manual 93-62 (14). Sandia Corporation. June 1962.

SCALES, E. M., & CHAPANIS, A. The effect on performance of tilting the toll-operator's keyset. *Journal of Applied Psychoogy*, 1954, 38, 452–456.

SCHEFFÉ, H. A method for judging contrasts in the analysis of variance. *Biometrica*, 1933, 40, 87–104.

SWAIN, A. D. *A method for performing a human-factors reliability analysis*. Sandia Corporation Monograph (SCR-685). August 1963.

(Received September 19, 1966)

journal of applied psychology

VOL. 51, No. 5, Part 1 of 2 Parts OCTOBER 1967

THE MULTITRAIT-MULTIRATER APPROACH TO MEASURING MANAGERIAL JOB PERFORMANCE

EDWARD E. LAWLER, III [1]

Department of Administrative Sciences, Yale University

The advantages of using the multitrait-multirater approach to measuring managerial job performance are considered. It is pointed out that, with this approach, it is possible to determine the convergent and discriminant validity of ratings, and because of this considerably more information can be obtained about the meaning of the ratings than could be obtained if a single-rater or single-trait approach were used. Multitrait-multirater data gathered from a sample of managers are analyzed and the convergent and discriminant validity of the ratings is determined. Encouraging levels of convergent and discriminant validity are obtained indicating that ratings potentially can be valid measures of managerial job performance. In addition, several off-quadrant analyses are performed that indicate looking at the disagreement among raters can lead to further understanding of the rating process. It is concluded that this approach has advantages for establishing criteria where they are needed, either for research purposes or for personnel decision-making purposes.

It has become increasingly stylish in the last 10 years to be critical of industrial psychology's favorite measure of job performance, the superior's global rating of his subordinate's performance. It is often pointed out that more objective measures are needed and that it is time to become more sophisticated in performance measurement. However, despite impressive pleas for new approaches, the superior's evaluation is still the most frequently used measure where criteria are needed either for research purposes or for personnel decision-making purposes. That is, whether a criterion is needed against which to validate a test, measure the impact of a training program, or whether a criterion is needed upon which to base a promotion or a pay raise, the superior's rating is still the most frequently used measure of job performance. For example, in Vroom's (1964) review of the studies relating job satisfaction and job performance, he found that 16 studies had used ratings while 7 had used objective measures of performance. A review of the criteria used for test validation as cited by Guion (1965) indicates that superiors' ratings are used about twice as frequently as more objective measures and that peer ratings are seldom used.

The superior's evaluation has probably enjoyed its greatest popularity at the management level. The reasons for this are obvious: management jobs are often multidimensional and hard to define; thus performance in them is difficult to quantify and make objective. This is not to say that attempts to get more objective criteria have not been made. Hulin (1962) and others (e.g., Bingham & Davis, 1924; Gifford, 1928; Williams & Harrel, 1964) have used salary as a criterion while others (e.g., Henry, 1948; Starch, 1942) have used organizational level achieved as a criterion. The problems with these indicators are in most cases even more severe than those associated with superiors' ratings. Salary, for instance, is easily quantifiable, but it is a result of a subjective rating to begin with, and then is further contaminated by the economic market value of managerial skills and a num-

[1] Revised version of a paper presented at the Executive Study Conference, Sterling Forest, N. Y., May 1966. The author would like to thank C. Argyris and L. W. Porter for their helpful comments on an earlier version of this paper.

ber of company policies. Thus, it is not surprising to note that several studies have concluded that salary level in management jobs is not necessarily closely related to merit (Lawler, 1966a; Meyer, Kay, & French, 1965). The same severe problems that exist for salary as a criterion are equally relevant with respect to promotion rate. As Stark (1959) has pointed out, promotion is a subjective decision that is made on the basis of factors that are often irrelevant as far as managerial effectiveness is concerned.

Organizational decentralization offers the opportunity in many cases for companies to establish profit centers against which to evaluate a manager's performance. This method does appear to have promise; however, it will probably never be possible to evaluate all managers in an organization on the basis of their own profit center. In addition, there is some evidence that unless profit-center measures and other objective measures are taken over long periods of time, they can be highly unreliable. Also, there is always a very real question as to whether profit and other objective measures are an accurate measure of the goals that managers should try to achieve in order to perform their jobs effectively. Further, many objective and seemingly relevant criteria may be completely beyond the control of the manager.

In-basket and other business games also have been suggested as potential criteria. They do possess some face validity and may be potentially very useful. However, there has been relatively little research on the relationship between how a manager behaves in a game and his behavior in an actual decision situation. The behavior of a manager in a business game may be quite different from his behavior on the job where the rewards and punishments are much larger. Further, business games tend to deemphasize the interpersonal dimension in managerial performance, while many management jobs appear to emphasize it heavily.

The multitrait-multirater approach to measuring performance has received relatively little attention, but it appears to be potentially quite valuable since it has some of the advantages of the more objective measures and some of those of the more subjective ones. For example, with this approach it is possible to assess the criterion by determining its convergent and discriminant validity, and it is not necessary to depend on an objective indicator such as profits or sales that may miss the essence of the job. The focus of the present paper, therefore, will be upon considering the potential usefulness of this approach both where research criteria are needed and where criteria are needed for personnel decision making. In addition, some empirical data will be presented relative to the usefulness of this method.

The first questions that arise when the multitrait-multirater approach is used concern who is going to participate in the multirater aspects of the approach and what traits are going to be included in the multitrait aspect. One must look first at the potential raters that are available for evaluating managerial job performance. The obvious constraint here is that the rating should not be done by ratees who are unfamiliar with the aspects of the individual's performance that they are to rate. Otherwise, the ratings tend to be more likely to be affected by the halo tendency (Bescoe & Lawshe, 1959) and tend to be unreliable. Thus, the most likely raters for a given manager would be his superior, his peers, his subordinates, and the manager himself. A good argument can be made that each of these raters typically has an adequate view of the manager's performance, although admittedly a slightly different one.

Superior ratings have traditionally been included because it is assumed that the superior has the best overview of the situation and knows best how the manager's job behavior contributes to the overall goals of the organization.

Peer evaluations are relevant because peers are best situated to evaluate how a manager performs in terms of the lateral relationships in working toward organization goals. Further, peers often see the manager at times when his superior is not viewing his behavior and, therefore, they may see aspects of his behavior of which the superior is not aware.

Subordinate ratings are relevant since subordinates are able to determine the superior's impact on what has been called by Likert (1961) the human resources of the organization. The subordinate is also often in a position to observe more of his superior's behavior than are either peers or superiors.

Self-ratings are relevant because the individual's self-perceptions are important determinants of his future behavior, in addition to the fact that he probably has more information about his own behavior than anyone else.

There are many potential traits that could be used in the multitrait aspect of the approach. Unfortunately, there is no easy rule that allows one to state that a certain list of traits can always be rated reliably and validly. The evidence seems to indicate that it is easy to err on the side of providing too many traits upon which to make ratings. On the other hand, as Dunnette (1963a, 1963b) and others have pointed out, the concept of a single criterion is unrealistic, and job performance is a function of a number of dimensions so that one needs to think in terms of partial criteria. The evidence from factor-analytic studies (e.g., Ewart, Seashore, & Tiffin, 1941; Grant, 1955; Rush, 1953) indicates that somewhere around three to five factors is a reasonable number. The question still remains, however, as to the kinds of traits upon which ratings should be obtained. One rating that probably should be included is a global one on quality of job performance. This statement is based partially upon Whitlock's (1963)[2] impressive series of studies on the psychological basis of performance judgments. His data suggest that when people are asked to make global ratings they act in a very predictable way, as efficient processors of critical-incident data from their observations of the individual's performance over the past 6 months. As Whitlock points out, judged quality of performance grows as a power function of the ratio of the number of specimens of effective to ineffective performance. Other studies have shown that raters tend to agree upon the weight to be assigned to the different behavior specimens; thus, interrater reliability is possible. This suggests that simple global-performance ratings may yield a reasonable approximation of what would be obtained by using a more extensive critical-incident or other type of checklist. The factor-analytic studies of ratings have also invariably produced a job-performance factor. In addition, Hollingworth's (1922) data on what traits can be rated reliably suggest that performance dimensions can be reliably rated.

A second reason for including a global rating is that it may help to reduce the halo effect that often comes from overall performance when other ratings are given. The assumption is that including the global rating allows raters to get their overall positive feelings toward the ratee out of their systems.

The question of what other traits should be included in addition to a global one is difficult to answer. The answer depends on the purpose of the study and on the particular types of behavior that characterize the important functions of the job. It may be necessary to develop partial criteria that are ratings of traits or factors like supervising, planning, and others which Hemphill's (1960) studies seem to indicate are the functions that make up a managerial job. The traits that are included should be clearly defined and carefully distinguished from the global-performance measure and from each other. Perhaps one operational way of determining whether a trait can be specified exactly enough is to see if a behavior-description-anchored rating scale can be developed for it. Barrett, Taylor, Parker, and Martens (1958) have found that formats incorporating behavioral description of scale steps were of superior reliability to numerically anchored scales. One test of how adequately a trait can be observed and conceptualized may be whether a behavior description can be developed for the scale points. Smith and Kendall (1963) have used behavioral descriptions provided by employees as anchors for rating scales and they report that excellent discrimination and high scale reliability were obtained. Further evidence in support of adding verbal description to scale points comes from a study by Peters and McCormick (1966) which showed that job-task-anchored rating scales are more reliable than numerically anchored scales.

One important perspective that has been too often overlooked where multitrait ratings systems are created is the development first of a theoretical point of view about what the

[2] G. H. Whitlock, Performance Evaluation through Psychophysics. Unpublished paper, University of Tennessee, 1965.

determinants of effective performance are. This does not mean a theory of whether performance consists of quality, productivity, turnover, or absenteeism, but rather attention to a model like the one that specifies that performance = f(Ability × Motivation). This model points out that job performance has two components—ability (a) and motivation (m)—and suggests that unless separate measures of these are obtained, errors are likely to appear in evaluating the effectiveness of ability tests and in evaluating the motivational systems of organizations. For example, a study that measures the effects of a pay plan by including only measures of performance (e.g., output, quality) would appear to be in danger of missing the impact of the pay program. Output measures are affected by ability as well as situational factors and they might mask the increased effort the workers are putting forth under the new plan. The real test of the plan should be whether or not it increases motivation.

A similar problem can occur in test validation. Ability tests are often evaluated against output and performance measures alone without including measures of motivation. A situation may occur where test scores are uncorrelated with performance measures because motivation is low. The model clearly points out that unless motivation is high, ability will not be related to performance. Thus, the wrong conclusion in this instance would be to assume that the ability measure is no good and to stop testing. A more realistic approach would be to keep testing despite low validity and to try to increase the motivational level of the employees so that the ability differences will manifest themselves. Since there is considerable evidence (Fleishman, 1958; French, 1957; Lawler, 1966b; Vroom, 1960) that indicates the performance = f(a × m) model has general validity, this would appear to be a good place to start. Specifically, it would appear to be reasonable to get ratings on a few factors like effort and ability that are suggested by the model and to avoid the traditional approach of giving a long list of traits including such dimensions as adaptability, friendliness, consideration, etc., which cannot be rated reliably (Hollingworth, 1922). Some support for the belief that the ability factor is a reasonable one upon which to obtain ratings comes from the Ewart, Seashore, and Tiffin (1941) study where an ability factor appeared in their factor analysis of rating scales.

Use for Research Criteria

At this point it seems appropriate to turn to the question of what the researcher can gain from using the multitrait-multirater approach, and to present some data concerned with the use of this approach in establishing research criteria. The primary gain from a research point of view is that this approach allows the researcher to develop a much more sophisticated understanding of his criteria than is possible where it is not employed. This understanding can come about partially as a result of being able to determine the discriminant and convergent validity (Campbell & Fiske, 1959) of the ratings.[3]

Convergent validity is demonstrated by the correlations between the same traits as rated by different raters being significantly different from zero. Discriminant validity is demonstrated by three criteria. First, a validity diagonal correlation value should be higher than the values lying in its column and row in the heterotrait-heterorater triangles. That is, a trait measure should correlate more highly with another measure of the same trait than with any other variable having neither trait nor rater in common. Second, a trait measure should correlate higher with an independent effort to measure the same trait than with measures designed to get at different traits which employ the same rater. For a given variable, this involves comparing its values in the validity diagonal with its values in the heterotrait-monorater triangles. Third, it is desirable for the same pattern of trait interrelationships to be shown in all of the heterotrait triangles of both the monorater and the heterorater blocks.

[3] Campbell and Fiske (1959) consider the multitrait-multimethod approach rather than the multitrait-multirater approach; however, they point out that use of raters that occupy different organizational positions relative to the ratee can reasonably be considered to be measurement by more than one method.

Campbell and Fiske (1959) point out that most studies do not report criterion measures that even begin to approach all the requirements for convergent and discriminant validity. The major problem, however, at this stage in the development of criterion measures for managerial positions, is not so much that ratings often fail to meet the requirements for either convergent or discriminant validity, but that in most studies presented in the literature, it is impossible to assess the convergent and discriminant validity of the criteria used. Given this absence of information, it is impossible to determine if one set of raters, or perhaps one trait, is performing particularly well, and therefore should be the central focus of the study. It is also difficult to eliminate certain traits or raters in order to sharpen criterion measurement for future studies. In effect, where either one rater is used or where ratings are obtained on only one trait the validity of criteria must be taken on faith. If insignificant results are obtained, it is impossible to know whether the problem lies in the criterion or in the predictor.

It should also be pointed out at this time that it is possible for ratings to have convergent and discriminant validity and still not be what would be normally called valid measures of the dimension that is to be measured. For example, peer and superior rankings of a group of managers may agree perfectly so that convergent and discriminant validity are obtained. However, both the peers and superiors may be simply making the same incorrect inferences from observing the individual's behavior. A manager who is seldom in his office and rarely available for meetings may be seen as low in commitment to the organization, in comparison with others, by both his peers and superiors. However, it may be that, unknown to the others, he is spending all his waking hours working on an innovation that will significantly help the organization. This would suggest that despite the existence of convergent and discriminant validity for the ratings on commitment, the ratings would be invalid. However, it is probable that this kind of situation is sufficiently rare to permit placing considerable faith in ratings that obtain convergent and discriminant validity. Still, this example clearly illustrates the point that it is impossible to ever finally validate a criterion. All that can ever be done is to gain information about what it measures.

TABLE 1

CRITERION INTERCORRELATION FROM TUCKER, CLINE, AND SCHMITT (1967)

Traits	Superior ratings				Peer ratings			
	A_1	B_1	C_1	D_1	A_2	B_2	C_2	D_2
Superior								
A_1								
B_1	.83							
C_1	.79	.71						
D_1	.91	.86	.83					
Peers								
A_2	.16	.04	.13	.15				
B_2	.30	.27	.26	.29	.66			
C_2	.16	.07	.24	.15	.75	.59		
D_2	.18	.09	.18	.16	.94	.71	.80	

Note.—N = 79.

With the requirements of convergent and discriminant validity in mind, two examples where the multitrait-multirater approach has been used can be considered. The first is shown in Table 1 and was taken from an article by Tucker, Cline, and Schmitt (1967). They had superiors, peers, and subordinates rate research scientists on four traits. Unfortunately, they did not report the correlations for the subordinates' ratings, although they do mention that they did not agree with the peer and superior ratings. Table 1 shows that the peer and superior ratings do not tend to agree on the validity diagonal (circled correlations) highly enough to argue that convergent validity exists. In addition, the data do not appear to satisfy the requirements for discriminant validity either. What is evident from the correlations is that a large halo tendency

TABLE 2

INTERCORRELATIONS AMONG RATINGS

Traits	Superior			Peers			Self		
	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Superior									
A_1									
B_1	.53								
C_1	.56	.44							
Peers									
A_2	.65	.38	.40						
B_2	.42	.52	.30	.56					
C_2	.40	.31	.53	.56	.40				
Self									
A_3	.01	.01	.09	.01	.17	.10			
B_3	.03	.13	.03	.04	.09	.02	.43		
C_3	.06	.01	.30	.02	.01	.30	.40	.14	

Note.—$N = 113$.

exists and that the superiors and peers are seeing quite different things.

Admittedly, at this point the researchers are not on completely firm ground with respect to their criteria, but they can now make more sensible interpretations of their data than if they had used only one set of raters or obtained ratings on only one trait. For example, the variable that the researchers were interested in was creativity. Had they obtained ratings of only creativity, they might have erroneously assumed that this is what they were measuring and gone on to say that their data showed that certain variables were related to creativity. In the present case, because multitrait data were collected, it is possible to look further at the creativity measures and to determine that they probably reflect the general positive halo of the individual which may be influenced by his creativity.

The second example of the use of a multitrait-multirater approach comes from some data recently collected by the author from a group ($N = 113$) of middle- and top-level managers in a manufacturing organization. Superior, peer, and self-ratings were obtained on quality of job performance, ability to perform the job, and effort put forth on the job. The superiors and peers were asked to do rankings and these were converted to standardized score equivalents of percent position for the purposes of the data analysis. Table 2 presents the correlation matrix for these data. It shows that the superior and average peer ratings have good convergent validity and it shows that they meet two of the three criteria for discriminant validity. The validity diagonal is higher than the correlations found in the heterotrait-heterorater triangles (dotted lines) and the same pattern of trait interrelationship is shown in all of the heterotrait triangles, even though there are some differences in the general level of correlations involved. The third requirement for discriminant validity that the validity diagonal corre-

lations be higher than the values in the heterotrait-monorater triangles (solid lines) comes close to being met. This is a rather stringent requirement which as Gunderson and Nelson (1966) point out is seldom met by behavior-trait data. The fact that some evidence for convergent and discriminant validity has been obtained is a significant point since it suggests that ratings of managerial job performance can achieve a level of measurement that is aspired to, but infrequently obtained.

A recent study by Gunderson and Nelson (1966) also obtained evidence for peer and superior ratings on three traits that indicates that they obtained convergent, and to some extent discriminant, validity. Forehand (1963) has found that forced-choice ratings of innovative behavior show considerable promise with respect to convergent and discriminant validity but that ratings obtained on 7-point rating scales do not. Fiske and Cox (1960) also have found that ratings on certain traits can achieve promising levels of convergent and discriminant validity. Thus, it would appear that ratings with a good degree of discriminant and convergent validity might be generally obtainable where well-defined traits are used, where the raters have good knowledge of the managers' job performance, and where appropriate rating scales are used.

Less encouraging are the self-ratings' data presented in Table 2. These data offer little evidence of either convergent or discriminant validity, as the self-ratings appear to be relatively unrelated to the superior and peer ratings. At this point the researcher is in the position of not knowing how to interpret the self-ratings. One possibility is that they are reflecting different, but nevertheless valuable, views of the same traits on which superiors and peers agree. A less charitable view is that they are simply totally invalid. It is true that the self-ratings are ratings rather than rankings and that because of this they are generally inflated. However, they may still be valid despite the leniency tendency if the relative position of the managers stays the same. Kirchner (1966) has found some low but significant relationships between self and superior multitrait ratings; however, he did

TABLE 3

MEAN SCORES ON DEMOGRAPHIC AND NEED-DISSATISFACTION VARIABLES FOR THE TWO OFF-QUADRANT GROUPS AND THE ON-QUADRANT GROUP IN SUPERIOR- AND PEER-RATINGS COMPARISON

	Lower by peers than superior ($N = 33$)	Same by superior and peers ($N = 82$)	Higher by peers than superior ($N = 32$)
Demographic variables			
Time in present position (% longer than 3 yr.)	42.4	37.4	64.5
Time with company (% longer than 10 yr.)	39.4	47.6	59.4
Age (% older than 40)	36.4	56.1	71.9
Education (% with college degree)	66.7	60.0	53.1
Needs			
Security (1 question)	.5	1.3	1.5
Social (2 questions)	.4	.5	.8
Esteem (3 questions)	.5	.8	1.3
Autonomy (4 questions)	.8	1.1	1.2
Self-Actualization (3 questions)	1.0	1.4	1.7

not report enough data to allow the discriminant validity of the ratings to be determined.

The question of how subordinates' ratings are related to superiors', peers', and self-ratings, of course, cannot be answered from these data. At this point it would seem to be entirely appropriate to focus on this topic in an empirical investigation. A reasonable prediction is that they would probably correspond more closely to the superior and peer ratings than to the self-ratings.

The use of the multitrait-multirater approach offers another benefit to the researcher since it enables him to explore a potentially quite fruitful research area. With this approach it is possible to begin to answer questions such as: what are the factors associated with managers being evaluated differently by

TABLE 4

MEAN SCORES ON DEMOGRAPHIC AND NEED-DISSATISFACTION VARIABLES FOR THE TWO OFF-QUADRANT GROUPS AND FOR THE ON-QUADRANT GROUP IN SUPERIOR- TO SELF-RATINGS COMPARISON

	Lower by superior than self (N = 40)	Same by superior and self (N = 81)	Higher by superior than self (N = 33)
Demographic variables			
Time in present position (% longer than 3 yr.)	52.5	40.8	33.3
Time with company (% longer than 10 yr.)	47.5	48.1	42.4
Age (% older than 40)	60.0	54.3	45.5
Education (% with college degree)	50.0	67.9	57.6
Needs			
Security	1.3	1.3	.8
Social	.6	.6	.3
Esteem	1.1	1.0	.6
Autonomy	1.3	1.1	.8
Self-Actualization	1.6	1.4	1.2

their superiors, peers, and selves? This can be accomplished by looking at the off-quadrants in the relationship between the different raters' ratings and looking for moderator variables (Ghiselli, 1963). The data reported in Table 2, when combined with some additional data gathered from managers in four government and social welfare agencies, offer a sample of managers on whom to attempt this kind of analysis. These data are particularly appropriate since job-satisfaction measures and demographic data were also collected for each manager. For the purpose of this data analysis, it was decided to focus on the ratings of job performance. Two off-quadrant groups (a low-high group and a high-low group) and an on-quadrant group were established for the relationship between the superior and the peer ratings, for the relationship between the superior and the self-ratings, and for the relationship between the peer and the self-ratings.

Table 3 presents the mean scores for the two off-quadrant groups and for the one on-quadrant group where the comparison is between the superior and peer ratings. The data were gathered from questionnaires that each manager completed at a group meeting where the researcher explained that the data were being collected for a university research project and that their responses would be confidential. As can be seen in Table 3, they were asked to indicate their age, how long they had been in the same position, their education, and their seniority. Where possible, these data were checked with company records and typically proved to be correct.

Table 3 also presents mean dissatisfaction scores for the three groups in five need areas. The 13 items used here are those originally used by Porter (1961) and contain two parts. The manager is asked to indicate on a 7-point scale how much of a given factor there is now associated with his job, and then he is asked to indicate how much there should be associated with his position. The rationale is that the larger the difference between his feelings of how much there is and his feelings of how much there should be, the greater the dissatisfaction. Thus, the larger the numbers in Table 3, the greater the dissatisfaction.

Two trends appear in the data presented in Table 3. The first trend appears in the demographic data which show that the off-quadrant group which is rated lower on job performance by their peers than by their superiors is composed of the younger ($p < .01$), lower seniority ($p < .10$), more highly educated managers (*ns*) than is the group that is rated higher by peers than by superiors. Thus, the off-quadrants appear to contain two different but rather identifiable groups. What may be happening in these organizations is that peers, perhaps through a sense of rivalry, are downgrading the young, bright lights in the organization. This tendency undoubtedly has been accentuated in several of these organizations since they recently have made special efforts to revitalize by bringing in new managers. This finding suggests that when peer ratings are used in a situation such as the one existing in these organizations, an attempt should be made to consider the degree to which they

reflect rivalries and resistance to new managers.

The second trend that is apparent from Table 3 is related to the satisfaction data. Those managers who are rated higher by their superiors than by their peers are better satisfied across the need areas than are those managers who are rated higher by their peers than they are by their superiors ($p < .01$).[4] An indication of the reason for this finding can be found in the responses of the managers to the question of how much is there now, which was asked for each of the 13 need-satisfaction items. Analysis of these responses showed that those managers who were rated higher by their superiors than by their peers felt that they were getting more in terms of need satisfaction than were those rated higher by their peers. This, of course, would be expected in an organization, since the superiors control the opportunities that managers have for obtaining many kinds of need satisfaction. Apparently, being rated more highly by the boss than by one's peers is not bad because one still gets rewarded, but being rated highly by peers and not the boss is disturbing because not only does one not get rewarded, one is likely to think that he should.

Table 4 presents the data from the comparison between the self- and the superior ratings. First, with respect to the demographic variables, the off-quadrants appear to be characterized as being made up of the extremes on the time dimension. The higher-by-superior-than-self quadrant in comparison to the lower-by-superior-than-self is made up of the younger managers. (Time in position is the only significant difference, $p < .10$.) Again, what may be happening is that the superiors are relatively more favorable toward their younger than toward their older subordinates.

Table 4 also shows that there is a tendency for those managers who are rated lower by

[4] A sign test was used in a manner similar to that done by Porter (1962). For each question the mean score for the lower-by-peers-than-superior group was compared with the mean score for the equal group. An increase was considered a plus and a decrease was considered a minus. A similar comparison was made for the equal group versus the higher-by-peers-than-superior group. Out of 13 questions there were 22 pluses, 3 ties, and 1 minus.

TABLE 5

MEAN SCORES ON DEMOGRAPHIC AND NEED-DISSATISFACTION VARIABLES FOR THE TWO OFF-QUADRANT GROUPS AND FOR THE ON-QUADRANT GROUPS IN THE PEER- TO SELF-RATINGS COMPARISON

	Lower by peers than self ($N = 46$)	Same by peers and self ($N = 68$)	Higher by peers than self ($N = 56$)
Demographic variables			
Time in present position (% longer than 3 yr.)	56.5	35.3	38.9
Time with company (% longer than 10 yr.)	54.4	44.1	47.2
Age (% older than 40)	58.7	54.4	52.5
Education (% with college degree)	50.0	73.5	52.8
Needs			
Security	1.3	1.3	.8
Social	.7	.5	.4
Esteem	.9	1.0	.7
Autonomy	1.1	1.1	.9
Self-Actualization	1.6	1.3	1.1

their superiors than by themselves to be more highly dissatisfied than are those managers who are rated lower by themselves than by their superiors ($p < .05$). An analysis of the managers' responses to the two questions that made up the satisfaction measure shows that the difference between two off-quadrant groups comes largely on the managers' answers to the question concerned with how much they should get. Those managers who were rated higher by their superiors than by themselves tended to have lower expectations about how much of the reward they should get than did those managers who were rated higher by themselves than by their superiors.

Table 5 presents the data from the comparison between the peer and self-ratings. Not surprisingly, since the peer and superior ratings are substantially correlated, the results for the peer-self rating analysis are similar to those obtained for the superior-self rating comparison although none of the differences

on the demographic variables is significant. High satisfaction appears where self-ratings are lower than peer ratings, and low satisfaction appears where peer ratings are lower than self-ratings ($p < .05$). As was the case with the superior-self rating comparison, this came about because of the difference among the groups on the managers' expectation of what they should receive. Higher self-ratings were associated with higher expectations.

In summary, the data show that the demographic variables of age, seniority, and time in position do tend to identify the off-quadrant groups in this sample. Managers with less seniority tend to be rated higher by their superiors than their peers and to be rated higher by their peers and superiors than by themselves. Older managers with high seniority tend to be rated lower by their superiors than by their peers and to be rated lower by their peers and superiors than by themselves. The data also show consistent relationships between satisfaction and the managers' positions in the on- and off-quadrant groups. High dissatisfaction appears where managers are rated lower by superiors than by peers and where managers are rated higher by themselves than by their peers and superiors. Low dissatisfaction occurs where managers are rated higher by superiors than by peers and where managers are rated lower by themselves than by their peers and superiors. These differences between the on- and off-quadrant groups are important because they can contribute substantially to the researcher's understanding of the basis upon which the ratings are made. Thus, a good argument can be made that, in addition to looking at the convergent and discriminant validity of ratings, off-quadrant analyses should be performed.

Implications for Personnel Decision Making

So far the focus has been upon the potential gains to a researcher from using the multitrait-multirater approach. There are also some advantages in using it for personnel decision making. Just as the opportunity to establish the convergent and discriminant validity of ratings is important for the researcher, it is important for the decision maker. This approach allows the decision maker to gain a real grasp of the adequacy of the superiors' ratings that have traditionally been used as a basis for personnel decisions. The expectation, therefore, would be that the opportunity to see the degree of relationship among the different trait ratings and the different raters' ratings will add significantly to the decision maker's understanding of what the ratings indicate. The result of this increased understanding is likely to be that decisions will be of a higher quality than if just superiors' ratings are relied upon. Further, the decision maker now has the opportunity of doing off-quadrant analyses which should lead to an increase in his understanding of the factors influencing performance appraisal in his organization. Thus, the same gains that are likely to accrue to the researcher from using this approach are also likely to come to the personnel decision maker if he uses this approach.

There are several important additional advantages that the decision maker gains from this approach, that lie in the multirater aspect of the method. Personnel decision making is at best a rather complex set of trade offs and compromises, whether it involves promotion, raises, or dismissals. By obtaining subordinate, peer, and self-evaluations, many of the trade offs inevitably will become much clearer in several important areas, perhaps the most important of these being motivation. It is the stated policy of most organizations that they attempt to reward merit with pay raises and promotions; yet there is also evidence that indicates many managers do not believe that merit is heavily rewarded in their organizations (Adams, 1965; Lawler, 1966a). One of the problems here may be that organizations have not taken careful enough account of peer and subordinate opinion when they have dispensed rewards in the past. Every time an organization gives a reward, it communicates to the members something about the payoff matrix of the organization. Unless an organization knows the perceptions of all its members with respect to the other members of the organization, it can never be sure what it is communicating when it dispenses rewards.

There is one final reason why decision quality can be improved through the use of

more than one rater. This advantage accrues from the fact that an individual's peers and subordinates are often in a better position to judge his performance and potential for other jobs than is his superior. Thus, improved decision quality may be expected from the additional relevant evaluations offered by other observers. The evidence with respect to subordinate ratings suggests that they may be valid but the evidence is not extensive. Mann and Dent (1954) have found that subordinates could identify some characteristics of promotable supervisors, and Patinka (1962) reported that less than 10% of those superiors who were presented with their subordinates' ratings of them thought they were invalid.

A rather large body of literature has now accumulated on the ability of peer ratings to predict other criteria, much of it contributed by Hollander (1954, 1956, 1957, 1964, 1965), and the evidence is consistently favorable. It has been demonstrated that after a relatively short period of time reliable test-retest ratings are obtained. Williams and Leavitt (1947) have found that peer ratings were better predictors of long-term success in the Marine Corps than were superiors' ratings. Roadman (1964) has found that peer ratings have validity as predictors of promotion in IBM, and Weitz (1958) found that peer ratings of salesmen during training correlated about .40 with superiors' ratings once the salesmen began work. Hollander (1965) and others (e.g., Wherry & Fryer, 1949) have also established that peer ratings do not tend to become popularity contests.

Perhaps the most important problem with respect to use of peer ratings for personnel decision making is the problem of the research set versus the administrative set. Most of the studies that have been done utilized the research set in getting their positive results. A good argument can be made that if the rater knows it is "going to count," ratings may lose their validity, particularly if a situation exists where an individual's self-interests might be best served by distortion of the peer ratings. However, a study by Hollander (1957) seems to indicate that peer ratings can be reliable and valid even when given with an administrative set. Undeniably, this is a topic that needs further research before wide use of peer ratings for administrative purposes would be advisable.

Leavitt (1964) has pointed out that peer ratings (the same would also appear to be true for self- and subordinate ratings) can be an important tool for allowing participation in personnel decision making, but that they will be effective only when integrated with a participative mood and participative practices in other areas. This relationship is illustrated and elaborated upon in Figure 1 It shows that the validity and reliability of a rating system for personnel decision making is not a simple function of the objective characteristics of the system, such as who does the rating or what characteristics are rated, but that individual differences as well as organizational characteristics modify this relationship. Under individual differences two dimensions appear to be important in determining reactions to rating systems. The first dimension concerns the strength of the need for feedback and the second is the authoritarian personality dimension. Presumably the more need an individual has for feedback and the less authoritarian he is, the more acceptable will be peer and subordinate ratings.

The assumption, with respect to organizational characteristics, is that in a Theory Y or participative organization, peer, self-, and subordinate ratings will be more acceptable than they will be in a Theory X organization. Attitudes toward the equity and acceptability of a rating system then are functions not only of the system itself but also of organizational and individual characteristics. The final part of the figure shows that the attitudes toward the system will dictate the validity of the ratings obtained. It is assumed that where negative attitudes exist among the raters toward the system, it will be impossible to get valid ratings. Perhaps the key point illustrated in Figure 1 is that a rating system does not exist in a vacuum and because of this no one system will be appropriate for all situations.

If one assumes for a moment that it is possible to establish the kind of organization in which peer ratings and perhaps even self-ratings can be valid, then an off-quadrant

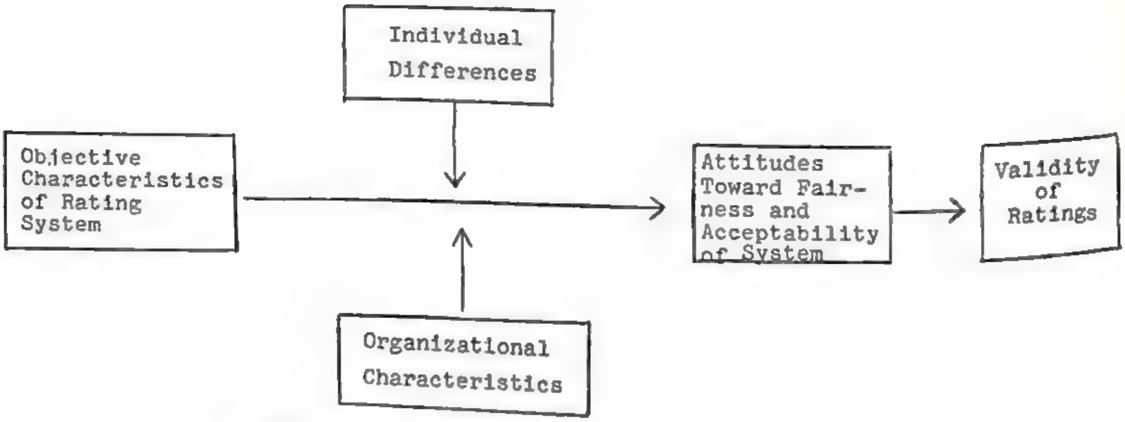


FIG. 1. Illustration of the factors that affect the validity of ratings.

analysis approach would appear to be particularly useful. Such an analysis would identify those managers about whom consensus does not exist and then these appraisal differences could be openly considered. For example, if a subordinate evaluates himself more highly than does his superior, he could openly discuss this with his superior and have the opportunity to point out to the superior what data he might be overlooking or perhaps misinterpreting. The results of this kind of interaction might well be a situation where few off-quadrant individuals can be found, but more importantly it could lead to a situation where each individual has more confidence in the evaluation situation because he has not only a feeling for how it operates but also has a chance to participate in it.

REFERENCES

ADAMS, J. S. *A study of the exempt salary program.* Crotonville, N. Y.: General Electric, Behavioral Research Service, 1965.

BARRETT, R. S., TAYLOR, E. K., PARKER, J. W., & MARTENS, L. Rating scale content: I. Scale information and supervisory ratings. *Personnel Psychology*, 1958, **11**, 333–346.

BESCOE, R. O., & LAWSHE, C. H. Foremen leadership as perceived by superiors and subordinates. *Personnel Psychology*, 1959, **12**, 573–582.

BINGHAM, W. V., & DAVIS, W. T. Intelligence test scores and business success. *Journal of Applied Psychology*, 1924, **8**, 1–22.

CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, **56**, 81–105.

DUNNETTE, M. D. A modified model for test validation and selection research. *Journal of Applied Psychology*, 1963, **5**, 317–323. (a)

DUNNETTE, M. D. A note on the criterion. *Journal of Applied Psychology*, 1963, **47**, 251–254. (b)

EWART, E. S., SEASHORE, S. E., & TIFFIN, J. A factor analysis of an industrial merit rating scale. *Journal of Applied Psychology*, 1941, **25**, 481–486.

FISKE, D. W., & COX, J. A., JR. The consistency of ratings by peers. *Journal of Applied Psychology*, 1960, **44**, 11–17.

FLEISHMAN, E. A. A relationship between incentive motivation and ability level in psychomotor performance. *Journal of Experimental Psychology*, 1958, **56**, 78–81.

FOREHAND, G. A. Assessments of innovative behavior: Partial criteria for the assessment of executive performance. *Journal of Applied Psychology*, 1963, **47**, 206–213.

FRENCH, E. G. Effects of interaction of achievement, motivation and intelligence on problem-solving success. *American Psychologist*, 1957, **12**, 339–400. (Abstract)

GHISELLI, E. E. Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, 1963, **47**, 81–86.

GIFFORD, W. W. Does business want scholars? *Harpers Magazine*, 1928, **156**, 669–674.

GRANT, D. L. A factor analysis of managers' ratings. *Journal of Applied Psychology*, 1955, **39**, 283–286.

GUION, R. M. *Personnel testing.* New York: McGraw-Hill, 1965.

GUNDERSON, E. K. E., & NELSON, P. D. Criterion measures for extremely isolated groups. *Personnel Psychology*, 1966, **19**, 67–80.

HEMPHILL, J. K. *Dimensions of executive positions.* Research Monograph No. 98, Bureau of Business Research, Ohio State University, 1960.

HENRY, W. E. Executive personality and job success. *AMA Personnel Series*, 1948, No. 120.

HOLLANDER, E. P. Buddy ratings: Military research and industrial implications. *Personnel Psychology*, 1954, **7**, 385–393.

HOLLANDER, E. P. The friendship factor in peer nominations. *Personnel Psychology*, 1956, **9**, 425–444.

HOLLANDER, E. P. The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology*, 1957, **41**, 85–90.

HOLLANDER, E. P. *Leaders, groups, and influence.* New York: Oxford University Press, 1964.

HOLLANDER, E. P. Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology*, 1965, **49**, 434–438.

HOLLINGWORTH, H. L. *Judging human character.* New York: Appleton, 1922.

HULIN, C. L. The measurement of executive success. *Journal of Applied Psychology*, 1962, **5**, 303–306.

KIRCHNER, W. K. Relationships between supervisory and subordinate ratings for technical personnel. *Journal of Industrial Psychology*, 1966, **3**, 57–60.

LAWLER, E. E., III. Ability as a moderator of the relationship between job attitudes and job performance. *Personnel Psychology*, 1966, **19**, 153–164. (a)

LAWLER, E. E., III. Managers' attitudes toward how their pay is and should be determined. *Journal of Applied Psychology*, 1966, **50**, 273–279. (b)

LEAVITT, H. J. *Managerial psychology.* Chicago: University of Chicago Press, 1964.

LIKERT, R. *New patterns of management.* New York: McGraw-Hill, 1961.

MANN, F. C., & DENT, J. K. The supervisor: Member of two organizational families. *Harvard Business Review*, 1954, **32**, 103–112.

MEYER, H. H., KAY, E., & FRENCH, J. R. P. Split role in performance appraisal. *Harvard Business Review*, 1965, **43**, 123–129.

PATINKA, P. S. Rate your supervisor program. *Social Science Research Reports.* New York: Standard Oil Company, 1962.

PETERS, D. L., & MCCORMICK, E. J. Comparative reliability of numerically anchored versus job-task anchored rating scales. *Journal of Applied Psychology*, 1966, **1**, 92–96.

PORTER, L. W. A study of perceived need satisfactions in bottom and middle management jobs. *Journal of Applied Psychology*, 1961, **45**, 1–10.

PORTER, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *Journal of Applied Psychology*, 1962, **46**, 375–384.

ROADMAN, H. E. An industrial use of peer ratings. *Journal of Applied Psychology*, 1964, **48**, 211–214.

RUSH, C. H. A factorial study of sales criteria. *Personnel Psychology*, 1953, **6**, 9–24.

SMITH, P. C., & KENDALL, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, **47**, 149–157.

STARCH, D. An analysis of the careers of 150 executives. *Psychological Bulletin*, 1942, **39**, 435.

STARK, S. Research criteria of executive success. *Journal of Business*, 1959, **32**, 1–14.

TUCKER, M. F., CLINE, V. B., & SCHMITT, J. R. Prediction of creativity and other performance measures from biographical information among pharmaceutical scientists. *Journal of Applied Psychology*, 1967, **51**, 131–138.

VROOM, V. H. *Some personality determinants of the effects of participation.* Englewood Cliffs, N. J.: Prentice-Hall, 1960.

VROOM, V. H. *Work and motivation.* New York: Wiley, 1964.

WEITZ, J. Selecting supervisors with peer ratings. *Personnel Psychology*, 1958, **11**, 25–35.

WHERRY, R. J., & FRYER, D. H. Buddy ratings: Popularity contest or leadership criteria? *Personnel Psychology*, 1949, **2**, 147–159.

WHITLOCK, G. H. Application of the psychophysical law to performance evaluation. *Journal of Applied Psychology*, 1963, **1**, 15–23.

WILLIAMS, F. J., & HARRELL, T. W. Predicting success in business. *Journal of Applied Psychology*, 1964, **48**, 164–167.

WILLIAMS, S. B., & LEAVITT, H. J. Group opinion as a predictor of military leadership. *Journal of Consulting Psychology*, 1947, **11**, 283–291.

(Received October 3, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 382–386

A COMPOUND-ATTITUDE SCALING TECHNIQUE FOR THE MEASUREMENT OF PARTISANSHIP IN LABOR-MANAGEMENT ISSUES[1]

JOSEPH P. ZWEIG

Sir George Williams University

In considering the measurement of partisanship in labor-management issues, it is surprising to find how few valid instruments exist in view of the importance of both labor and management in an industrial society. Attention is directed in this article to a new scale which was originally constructed for the purpose of the measurement of partisanship attitudes of foremen. A novel feature of this instrument is a tailor-made neutral zone built into the scale which permits further analysis. It is possible that this technique may find applications elsewhere in psychology.

Measurement in the area of labor-management relations has engaged the attention of several investigators in recent years. England (1960) has summarized a number of studies on dual allegiance going back to 1949. In some of these studies the method of measurement used by the researchers consisted of interviewer ratings of favorableness and unfavorableness toward management and labor. This technique requires extensive, time-consuming interviewing of the subjects and subsequent analysis of the responses and is almost impossible to replicate.

Another method involved the use of two separate scales, the Union Attitude Scale and Employee Attitude Scale, developed at the Industrial Relations Center of the University of Minnesota. These scales were combined by England in the form of a questionnaire. While the items of the Union Attitude Scale may be regarded as measuring the degree of labor orientation, the items of the Employee Attitude Scale appear to measure satisfaction with the company as an employer and were developed as a "general indicator of employee morale" rather than as an indicator of the promanagement attitudes of the respondents. It would be difficult to identify or define by this device those cases which were neutral or "in the middle" in their preferences for siding with management and labor simultaneously. This difficulty is further increased by the fact that the zero or neutral score on the Likert-type scale which was used is ambiguous and capable of various psychological interpretations.

Bernberg, Cole, Giedt, Peters, and Weschler (1948, publ. in 1952) experimented with an indirect method, the error-choice technique, for measuring attitudes toward labor and management. In this type of scale the respondent is offered two choices both of which are in error. The choices are so designed that if there is a correct answer, the errors are made to fall on both sides of it. The principle involved is that the perception and recall of objective data will tend to be distorted by a respondent to suit his predispositions.

This work was subsequently refined by Wechsler (1952). Balma (1954) also developed a measure of management identification based on this technique. Though the error-choice technique offers the advantage of screening to prevent stereotyped responses, it has at least one drawback. The scores are inferences made not from the knowledge of the respondent but from his default. Furthermore, the items are usually composed of general informational material and therefore the scale also suffers from the difficulty of interpreting neutral scores from a psychological viewpoint.

It would seem reasonable that items intended to measure partisanship tendencies in the stormy area of management-labor relations should be characterized by a distinctly

---

[1] This scaling technique was devised by the author n connection with a larger study submitted to Teachers College, Columbia University in partial fulillment of the requirements for the degree of Doctor f Philosophy. The author wishes to acknowledge he guidance of his advisers, Albert S. Thompson, hairman of the committee, Robert L. Thorndike, nd Donald E. Super.

controversial flavor. To use items composed of informational material of a noncontroversial nature is to put greater emphasis on the cognitive component of the attitude when it should be on the emotional component.

In short, it appears that a need exists for a new kind of labor-management partisanship scale. Such a scale should tap the emotional component of the attitude. It should also permit the identification of the neutral zone of scores so that they can become amenable to psychological analysis.

PROCEDURE

Controversies in the field of labor-management relations originate as a rule in grievances which are drawn to the attention of management by some official representative of the work group. Many such grievances are in fact solved at the level of the supervisor and the shop steward, but some are not and go through a series of steps known as the grievance procedure. These are generally settled by an arbitrator whose decision is binding on both sides. The use of genuine grievances of this latter kind suggested the possibility of obtaining suitable items for a partisanship scale.

Seventy-two case histories which met the requirements for inclusion were assembled. In converting the case history to a suitable scale item, the core or essence of each issue was extracted, reworded, and each case was presented as a condensed problem in industrial relations.

Two possible solutions were framed and attached to each item as statement choices. One of these statements expressed a management point of view and the other a labor point of view. These viewpoints were matched and balanced as far as possible and were made to appear equally reasonable and justifiable but opposite in direction.

Two six-step Likert-type scales were then attached, one to each statement choice. In scoring each item the ratings on these two scales were combined. By this device it became possible not only to tell when the respondent favored or opposed one side or the other, but also the degree of his alignment and the manner of his nonalignment or neutrality. Thus if a subject did not side with management or with labor, it became possible to state where he stood and by what means he maintained his position.

The following is an example of a typical item extracted from the scale:

The Russet Works suspended operations for several days to permit the taking of the annual inventory. Several junior men were kept on for inventory check-up, but senior men were laid off. The seniors complained.

(a) A company has the right to take inventory at lowest cost to itself and since the work involved is not directly related to production it is not obliged to follow the rules of seniority.

+3	+2	+1	−1	−2	−3
agree strongly	agree moderately	agree more than disagree	disagree more than agree	disagree moderately	disagree strongly

(b) Any shut down which results in the release of workers from work is a lay-off and under such circumstances the company should follow seniority rules.

+3	+2	+1	−1	−2	−3
agree strongly	agree moderately	agree more than disagree	disagree more than agree	disagree moderately	disagree strongly

To establish construct validity the items were submitted to a team of five raters. By this means 21 items were found to be unacceptable leaving 51 items. These 51 items were then arranged in the form of a pretest to provide the data for the item analysis.

Item Analysis

The governing features of the item analysis are included in the following steps:

(*a*) Submit the pretest to criterion groups of known top-level managers and labor leaders.

(*b*) Eliminate those items on which these criterion groups agree.

(*c*) Select those items on which the criterion groups disagree in such a way as to maximize the disagreement on either side of a central value or position of neutrality.

The item analysis brought the number of items judged acceptable to 20. In the final step the significance of the difference between percentages of managers and labor leaders who gave promanagement responses was determined on these 20 items and the procedure was repeated for the prolabor responses. All of these differences were found to be significant at better than the .01 level. A comparison of the characteristics of the two groups on the 20 best items appears in Table 1.

Validity

The items of which the instrument was composed are essentially drawn from actual industrial grievances. In this sense the scale possessed representative, rational, or content validity. The statement choices attached to each item were judged by a team of raters to favor either labor or management

TABLE 1

COMPARISON OF SCORES OF 60 MANAGERS AND 50 UNION OFFICIALS ON THE 20 BEST ITEMS OF THE PARTISANSHIP SCALE, BY PARTISANSHIP, DUAL AGREEMENT, AND DUAL DISAGREEMENT

	N	$\bar{X}$	*SD*	Range
Partisanship				
Managers	60	57.8	18.9	23 to 113
Union officials	50	−52.6	23.4	−94 to 3
Dual agreement				
Managers	60	12.1	8.7	0 to 36
Union officials	50	13.5	10.6	0 to 42
Dual disagreement[a]				
Managers	60	2.1	3.3	0 to 8
Union officials	50	2.6	3.6	0 to 16

[a] All dual-disagreement scores were negative and the signs of these scores have been reversed to show absolute values.

permitting the acceptance of one side and the rejection of the other, the acceptance of both sides, or the rejection of both sides. These acts signified unilateral alignment, dual agreement, or dual disagreement, respectively. These results were further sharpened by item analysis using known contrasting groups. Consequently, the instrument may be said to have construct validity.

Finally, in order to establish statistical validity, a criterion was developed independently by the use of the Remmers-Kelly (Kelly & Remmers, 1934) "Scale for measuring attitude toward any institution Form B."[2] For this purpose a measure of attitudes toward "Big Business" was obtained in order to find the degree to which the respondents favored Management and a measure of attitudes toward "Labor Unions" was obtained in order to find the degree to which the respondents favored Labor. It was considered reasonable that differences between these institutional attitudes obtained from the same individuals should represent partisanship tendencies and should correlate positively with other measures of partisanship.

Two copies of the Remmers-Kelly Scale, labeled as indicated and alternated to avoid response set, were administered anonymously to a group of 78 Commerce students in the evening division of Sir George Williams University, Montreal. Concurrently the 20-item partisanship scale described in this study was also administered to these students. The scores of these students are shown in Table 2 and the validity coefficients are presented in Table 3.

The validity coefficient of greatest consequence is the last one mentioned on Table 3, namely, Partisanship and Difference Scores, $r = .64$.

Reliability

Reliability coefficients of the partisanship, dual-agreement, and dual-disagreement scores were computed from the data obtained for the 78 Commerce students. Additional reliability measures were obtained from a tryout sample of 35 managers. The coefficients are presented in Table 4.

Scoring Procedure

The scoring system yielded three scores for each respondent, a partisanship score, a dual-agreement score, and a dual-disagreement score. The partisanship score represents the degree of unilateral alignment. The other two scores represent literally the extent to which the respondent agrees or disagrees simultaneously with both sides on any of the items.

For convenience the scoring system was set up so that responses scored as promanagement were positive while those scored as prolabor were negative. To obtain the partisanship score in any item the labor-statement choice score is subtracted algebraically from the management-statement choice score. These item scores are then summed.

The dual-agreement and the dual-disagreement scores are each arrived at separately by merely summing the positive pairs and the negative pairs in those items in which the respondent indicates the one or the other. Since the dual-agreement scores must all be positive and the dual-disagreement scores must all be negative, there can be no confusion from this procedure.

There is no zero point on the scales attached to each of the statement choices thereby obliging the respondent to take a stand on each issue. In combining the two statement choices, however, a zero partisanship score results from the summing of equal or nearly equal but opposite scores.

In the case where a subject answers consistently, his score will show high partisanship, depending on whether he favors labor or management. Otherwise it will show reduced partisanship with or without dual agreement or disagreement. Since each item contributes a maximum of 6 points to the total score, for 20 items the maximum possible score is 120 points plus or minus, making a total possible range of 240 points. This is ample provision for a variety of response combinations.

TABLE 2

PARTISANSHIP, AGREEMENT, AND INSTITUTIONAL ATTITUDE OF 78 COMMERCE STUDENTS AT SIR GEORGE WILLIAMS UNIVERSITY

	N	$\bar{X}$	*SD*	Range
Partisanship	78	27.4	26.4	−58 to 100
Dual agreement	78	17.0	14.9	0 to 97
Dual disagreement[a]	78	2.8	3.6	0 to 14
Pro "Big Business"	78	8.7	1.8	3.6 to 10.2
Pro "Labor Unions"	78	7.1	2.5	2.5 to 10.2
Difference scores[b]	78	1.7*	3.1	−6 to 7.4

[a] All dual-disagreement scores were negative and the signs of these scores have been reversed to show absolute values.
[b] Difference scores were obtained by subtracting the "Labor Unions" scores from the "Big Business" scores.
* $p > .001$.

[2] By permission of H. H. Remmers, Division of Educational Reference, Purdue University, Lafayette, Indiana, personal communication, July 13, 1962.

TABLE 3

VALIDITY COEFFICIENTS OF THE PARTISANSHIP SCALE

Correlates	N	Pearson's r
Partisanship and "Big Business"	78	.39*
Partisanship and "Labor Unions"	78	−.54*
Partisanship and Difference Scores	78	.64*

* $p > .001$.

The matrix in Figure 1 indicates all the combinations of partisanship and agreement responses possible on any single item, along with the corresponding scores.

Partisanship

	+3	+2	+1	-1	-2	-3
-3	6	5	4	2	1	0
-2	5	4	3	1	0	-1
-1	4	3	2	0	-1	-2
+1	2	1	0	-2	-3	-4
+2	1	0	-1	-3	-4	-5
+3	0	-1	-2	-4	-5	-6

Agreement

	+3	+2	+1	-1	-2	-3
-3				-4	-5	-6
-2				-3	-4	-5
-1				-2	-3	-4
+1	4	3	2			
+2	5	4	3			
+3	6	5	4			

FIG. 1. Scoring matrix.

DISCUSSION

The compound scaling technique described here provides opportunities for a greater variety of responses than the customary single scale while preserving the general features of such scales.

By using two six-step Likert-type scales in combination to form an attitude continuum, the following responses can be distinguished in each item assuming that A and B are opposite ends of the continuum:

(*a*) Favor A and oppose B: Siding with A.
(*b*) Favor B and oppose A: Siding with B.
(*c*) Favor A and B simultaneously: Dual agreement.
(*d*) Oppose A and B simultaneously: Dual disagreement.

The combination of these responses resulting from sometimes siding with labor and sometimes with management in varying degree, and from different amounts of agreement and disagreement, allows considerable flexibility in the nature and number of global distinctions which can be made.

One of the advantages of the compound-attitude scaling technique is that it makes available three scores from the same set of responses. In this case they were a partisanship score, a dual-agreement score, and a dual-disagreement score. Another advantage is that it allows a tailor-made hidden neutral attitude zone to be built into the scale during its construction while obliging the respondent to take a stand on the issues in question. With the type of item analysis used here it is possible to align this neutral zone with the zero of the scale.

The neutral zone can be identified by first finding the standard error of measurement of an individual score from a knowledge of the standard deviation and the reliability of a sample of partisanship scores. Next, using the zero of the partisanship scale as a working origin, the interval containing two standard errors of measurement is determined on both sides of zero. This succeeds in establishing a medial interval lying half way between the extremities of the scale as the theoretical neutral zone. An individual's score must fall within this medial interval over 95% of the time, if his true score is taken to be zero on the partisanship scale.

Whether the zero on the partisanship scale is the best working origin to use as an anchor for the neutral zone is a matter which has to be taken care of in the construction of the scale. In practice any central value which lies midway between the actual mean partisanship scores of the labor and management criterion groups can be used with justification. If the statement choices in each item have been well matched and well balanced then this

TABLE 4

RELIABILITY COEFFICIENTS OF THE PARTISANSHIP SCALE

	N	r	Method
Partisanship			
Commerce students	78	.74	odd-even corrected[a]
Commerce students	55	.84	test-retest after 6 wk.
Managers (Tryout group)	35	.33	odd-even corrected
Dual agreement			
Commerce students	55	.51	test-retest after 6 wk.
Managers (Tryout group)	35	.64	odd-even corrected
Dual disagreement[b]			
Commerce students	78	.42	odd-even corrected
Managers (Tryout group)	35	.36	odd-even corrected

[a] Split-half reliability coefficients corrected by the Spearman-Brown prophecy formula.
[b] The dual-disagreement distribution was violently skewed and the reliability coefficient is open to question since the arrays did not appear to possess homoscedasticity.

central value should fall very close to, if not actually right on, the scale zero position.

With the medial interval, linked to a central value or actual zero, serving as a criterion, it becomes a simple matter to distinguish those individuals who are "in the middle" or neutral from those who exhibit promanagement or prolabor partisanship. When the neutral attitude zone is defined in this way the medial cases which fall within the zone become available for further examination.

In the major study conducted on first-line supervisors by Zweig (1964), from which this partisanship scale was extracted, it was possible to separate a medial group of 85 foremen into four more or less distinct subgroups classified as High Dual Agreement, High Dual Disagreement, Weak Partisan, and Nonpartisan. Only the Weak Partisan group could reasonably be characterized as neutral in that it tended to be impartial, lukewarm, or halfhearted in its support of one of the sides. The remaining three subgroups achieved mediality in a manner which, from a psychological viewpoint, could hardly be characterized as neutrality.

The High Dual Agreement subgroup achieved mediality through what appeared to be inconsistency since such cases involve simultaneous agreement with statement choices which were designed to represent opposite points of view. The High Dual Disagreement group achieved mediality through what may have been ignorance of the problems or dissatisfaction with the solutions offered in the statement choices or possibly evasiveness. The Nonpartisan group achieved mediality through what seemed to be confused alignment since they tended to give promanagement responses for some of the items and prolabor responses for the remainder.

What appears to become evident is that the interpretation of the middle position on an attitude scale to mean impartiality or neutrality is not only superficial but may in fact be incorrect. When the investigation of medial cases is made possible by exposing the neutral zone as defined above, several other ways of attaining mediality may emerge which have little to do with impartiality or neutrality. This kind of analysis is not readily available through conventional scales.

The partisanship scale also appears to be able to discriminate between such diverse groups as top management, foremen, labor officials, and commerce students. The final instrument was found to be very acceptable in use even though it dealt with controversial issues. Although no time limit was set, it rarely took over 20 minutes to complete.

The reliability coefficients shown on Table 4 are not all that may be desired, but they may be enough for group prediction. Although further validation is necessary, the present validity coefficient is encouraging. The validation suggests that the partisanship variable tapped by the scale is the residual of the difference between a subject's actual prolabor and his promanagement attitudes. The use of items composed of controversial issues has grounded the partisanship attitude in the subject's attachment for one side or the other, rather than in his belief system which may reflect tendencies toward conventionality.

The scale described above was designed for research purposes in industry. However, the technique is flexible, brief, and direct, and there is no reason why it may not be applied for research or practical purposes in a variety of areas where opposing views exist on the same issue.

REFERENCES

BALMA, M. J. Development of a measure of management identification using the error-choice technique. Unpublished doctoral dissertation, Purdue University, 1954.

BERNBERG, R., COLE, E., GIEDT, F. H., PETERS, D., & WESCHLER, I. R. A preliminary development of a labor relations attitude scale by the error-choice method. Unpublished 1948. In M. L. Blum (Ed.), *Readings in experimental industrial psychology.* New York: Prentice-Hall, 1952. Pp. 140, 147.

ENGLAND, G. W. Dual allegiance to company and union. *Personnel Administration,* 1960, *23,* 20–25.

KELLY, I. B., & REMMERS, H. H. A scale for measuring attitude toward any institution Form B. Lafayette, Indiana: Purdue Research Foundation, 1934.

WESCHLER, I. R. An investigation of attitudes toward labor and management by means of the error-choice method. In M. L. Blum (Ed.), *Readings in experimental industrial psychology.* New York: Prentice-Hall, 1952. Pp. 140–147.

ZWEIG, J. P. The relationships among performance, partisanship in labor-management issues, and leadership style of first line supervisors. Unpublished doctoral dissertation, Columbia University, 1964.

(Early publication received February 27, 1967)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 387–392

MEASUREMENT OF INDUSTRIAL RELATIONS ACTIVITIES[1]

PHILIP ASH

Inland Steel Company, Chicago, Illinois

43 major corporations provided data on organization structure and employment, and responses to a 38-item industrial relations activities scale. Wide variations were found both in the range of activities undertaken by the industrial relations department and in the extent to which responsibility for these activities is centralized at the corporate level or decentralized into the operating units of the company. Indexes of activity level and activity centralization seemed to be in agreement with reported policies and practices. There were consistent and significant differences with respect to centralization as among specific industrial relations activities (e.g., union contract negotiations are typically centralized, time and motion study is typically decentralized). A promising scale for the measurement of organizational centralization and activity level in the industrial relations functions has been developed.

While each company must shape its industrial relations function to suit its own particular local needs and situation, it is desirable from time to time to evaluate how this function is being discharged, either by comparison with other companies to determine the extent of divergence from the common pattern, or by manipulation of internal data without comparison.

Where intercompany or interindustry benchmarks are available the evaluation will provide a more realistic appraisal than would be possible in the absence of such benchmarks.

In recognition of the desirability of creating benchmarks for comparative purposes, the Industrial Relations Center of the University of Minnesota, under the direction of Dale Yoder, undertook an annual survey of salaries and staffing ratios in industrial relations. In all, between 1949 and 1962, when the project was abandoned, 13 annual surveys were published (Nelson, England, & Yoder, 1963). The central statistic developed was the personnel ratio (*PR*): the ratio of the number of employees in industrial relations to the total employment in the company. This ratio has been widely used in industry to assess the industrial relations function (Mee, 1958, p. 112). As Gray (1965) and others have pointed out, however, the *PR* provides less than unambiguous information. The ratio also may be correlated with company size (American Management Association, 1961, Table 12; Nelson, England, & Yoder, 1960; Wortman, 1963), limiting the possibility of applying an average industry figure to evaluate the industrial relations functions in different-sized companies.

Furthermore, the number of employees in a function such as industrial relations is not necessarily a very good indicator of what they do, considering the wide range of possible activities in the function.

A more sophisticated statistic to evaluate the industrial relations function (or any other corporate activity, for that matter) on the basis of comparative work-force data is a regression equation including the predicted number of employees in the function as the dependent variable, and total employment as the independent variable, namely,

$$IR = a + bE$$

where *IR* is the predicted number of employees in industrial relations, *E* is total employment, and *a* and *b* are the constants.

It may be noted that if the linear relationship does exist, it follows that the *PR* systematically varies with total company employment:

$$PR = \frac{IR \text{ employment}}{\text{Total employment}}$$

$$= \frac{a + bE}{E}$$

$$= \frac{a}{E} + b$$

[1] An expanded version of a paper presented to a meeting of the Midwestern Psychological Association, Chicago, April 1965.

The American Management Association (1961, Table 12) study of 14 industries, cited above, showed that a linear-regression line fits the data very well. The correlations between *IR* employment and total employment ranged from .73 to .99 for the 14 industries, and was .89 for all companies combined without respect to industry. In 7 of the industries, the intercept (*a*) was positive: the PR *decreased* with increasing company size. In the remaining 7 industries, however, the intercept was negative: the PR increased with increasing company size.

A third approach to evaluating the industrial relations function has involved the construction of an index based upon activity measures such as accession and separation rates, grievances, suggestions submitted, safety data, etc. (French, 1954; Jackson, 1961; Merrihue & Katzell, 1955). Typically, in such an index, six to eight measures of industrial relations activities are combined into a weighted composite to provide a single-number measure of the state of the function. At least in the published studies, one important limitation has been the lack of cross-company comparison and, indeed, the practical impossibility of making such a comparison, since the same functions, measured in the same ways, do not seem to occur across any large number of companies.

METHOD

The present study was an attempt to develop other measurement concepts and benchmarks which might contribute to understanding the ways in which industrial relations staff may be deployed to accomplish many different objectives, with particular emphasis on the conceptualization and measurement of level of activity and of centralization versus decentralization.

Forty-three[2] major corporations provided, via a lengthy questionnaire, data on organization structure and on general employment and industrial relations employment, and responses to a 38-item scale of industrial relations activities (listed in Table 3). The 43 companies included 13 steel companies ranging in employment size from 9,700 to 244,000 workers; the median was 16,000. The 30 other companies were distributed among 12 manufacturing industries and reported total employment ranging from 11,400 to 38,000 employees; the median employment was 30,000. For the sample as a whole, median employment was 26,950.

A correlation analysis was made of data on 11 variables, and measures of centralization were also compared with qualitative data on centralization policies and practices. The 11 variables included:

First, three measures of employment size:

(1) total employment,
(2) total industrial relations staff,
(3) headquarters or corporate industrial relations staff.

Total headquarters work force was also noted, but not treated as a separate variable.

Second, two traditional personnel ratios:

(4) total industrial relations personnel ratio (industrial relations staff to total employment), and
(5) corporate industrial relations personnel ratio (corporate staff to total employment).

Third, two measures of centralization of staff based upon employment figures:

(6) the ratio of corporate industrial relations staff to total industrial relations staff, and
(7) a personnel concentration index based on a comparison of the previous ratio with the ratio of total corporate staff to total work force.

This index measures the extent to which the industrial relations function is more or less concentrated at the corporate level than are all other company functions.

Fourth, two measures of industrial relations activity based on responses to the 38-item scale. For each item, the responding company indicated that (*a*) the corporate staff conducts and administers the activity (e.g., labor negotiations), (*b*) the corporate staff reviews and approves, (*c*) the corporate staff provides advice and counsel only, (*d*) the activity is not a corporate activity, or (*e*) there is no active program in the area in the company.

The two measures based on this score are:

(8) an activity centralization index, a score based on the first four response categories.

This score can vary from 0 (none of the listed activities is a corporate responsibility) to 3.00 (all activities centrally located, conducted, and administered), and

(9) an activity ratio—the number of activities coded to reflect corporate responsibility for at least advice and counsel (Codes a, b, or c checked), divided by 38.

Finally, two measures of the ratio of professional, technical, and managerial (exempt) staff to clerical (nonexempt) staff:

[2] Not all companies reported complete data. For each variable, the number of responding companies is shown in Table 1.

TABLE 1

MEDIANS, QUARTILES, MEANS, AND STANDARD DEVIATIONS FOR ALL INDUSTRIAL RELATIONS (IR) MEASURES

Measure	No. responding companies	*M*	*SD*	Q 1	*Mdn*	Q 3
1. Total employment	43	32826	40647	14801	26950	31548
2. Total IR staff employment	28	284	334	100	176	315
3. Corporate IR staff employment	37	35	37	13	26	41
4. Total IR personnel ratio	28	1.01	0.59	0.63	0.85	1.12
5. Corporate IR personnel ratio	37	0.11	0.10	0.05	0.09	0.16
6. Corporate IR staff to total IR staff	26	16.00	10.61	8.93	15.60	19.90
7. Personnel concentration index	26	5.12	3.66	2.66	4.55	6.97
8. Activity centralization index	42	1.65	0.53	1.26	1.72	1.93
9. Activity ratio	42	70.20	19.30	64.00	74.90	82.90
10. Exempt/Nonexempt ratio: Total IR staff	26	1.08	0.61	0.73	0.87	1.35
11. Exempt/Nonexempt ratio: Corporate IR staff	26	1.66	0.86	1.16	1.62	2.22

(10) the exempt-nonexempt ratio for the total industrial relations function, and
(11) the exempt-nonexempt ratio for the corporate industrial relations function.

RESULTS

Tables 1 and 2 summarize, respectively, measures of central tendency and dispersion on the 11 variables and the intercorrelations among the variables.

Employment and the personnel ratio. While the three employment-size statistics are highly correlated with one another (as might be expected), they are related to only a limited extent to the rest of the measures. The correlations of total employment with the various ratios and indexes are all insignificantly different from zero. It might be noted that the correlations between total employment and both *PR*s (total *PR* and corporate *PR*) are both negative. At least in this sample, as company employment size increases, the *PR* tends to decrease. The absolute size of the corporate industrial relations staff, however, is *positively* related to the corporate *PR*.

Total PR and centralization. The total *PR* is negatively correlated (−.40, −.38) with the two measures of centralization (personnel concentration index and activity centralization index). Decentralized companies, in other words, tend to have larger industrial rela-

TABLE 2

INTERCORRELATION MATRIX

Variable	2	3	4	5	6	7	8	9	10	11
1. Total employment	89	71	−08	−15	−10	18	12	03	−17	−16
2. Total IR staff employment	—	72	36	−03	−16	−00	−05	−02	−26	−25
3. Corporate IR staff employment			08	51	39	18	31	22	−19	−23
4. Total IR personnel ratio			—	−20	−06	−40	−38	−10	−30	−09
5. Corporate IR personnel ratio				—	71	02	20	25	04	−18
6. Corporate IR staff to total IR staff					—	25	44	50	08	00
7. Personnel concentration index						—	−04	−04	30	−21
8. Activity centralization index							—	53	06	30
9. Activity ratio								—	12	10
10. Exempt/Nonexempt: Total IR staff									—	35
11. Exempt/Nonexempt: Corporate IR staff										—

Note.—Decimal points omitted.

tions staffs in relation to their total employment than do companies in which the industrial relations function is more centralized at the corporate level.

Centralization and level of activity. Centralization of functional responsibility at the corporate level (activity centralization index) tends to be related to a wider range of industrial relations activities (activity ratio; $r = .53$), and both functional centralization (activity centralization index) and a wider range of activities (activity ratio) are associated with larger corporate staffs vis-à-vis industrial relations staffs ($r = .44$ and .50, respectively).

Exempt-nonexempt ratio. Contrary to original expectations, the ratio of exempt to nonexempt personnel did not seem to be related to any of the other indexes or measures that were computed. Organizations that have fairly well-centralized industrial relations programs are about as likely to have low exempt-nonexempt ratios as high exempt-nonexempt ratios, and decentralized programs reflect similar variety. It had been thought that as industrial relations programs increase in size, extent, and variety of functions, and degree of centralization, the ratio of exempt to nonexempt employees would go down because larger numbers of clerical people would be required to accomplish all the details of administration. This hypothesis is at least partially sustained in the comparison of the ratios for the total industrial relations force with the ratios for the corporate industrial relations force. The latter are consistently higher, reflecting the relatively larger number of clerical personnel involved in processing of forms and records at the plant level. However, there does not appear to be a consistent pattern across the companies surveyed for either of the ratios considered separately.

The activity centralization index and perception of centralization. For the sample as a whole, the index ranged from 0.00 to 2.69 with an overall median of 1.72 (the steel industry is characterized by a slightly higher tendency toward industrial relations centralization—median 1.83—than nonsteel companies—median 1.53).

To test the index against management *perceptions* of centralization, respondents were asked for their views concerning the extent to which industrial relations activities are, or should be, centralized at the company-headquarters level. Companies with high activity centralization indexes generally reported that the function *was* centralized, and the leading reason cited was the need for uniformity of action arising out of master labor agreements.

For example, a steel company with a score of 2.69 said

> . . . from our experience there is a great need for a close review and approval of the activities listed through a central staff functioning at Corporate or Company Headquarters. Unless this is done there can be misinterpretation of policies which would ultimately go far afield of the original purpose and intent of such policies. We have found this to be very definitely the case even where operations are closely integrated. It is very easy to create misunderstanding involving such matters as Employee Benefit Programs (insurance and pensions), through wrong interpretations at plant levels. Such matters can be corrected quickly where general clearance is established through centralized headquarters staff. . . . In certain areas such as those related to safety, the actual operations must be carried out at the plant and departmental levels. However, the great majority of the other activities depend primarily on strict interpretation of policies established by top management and therefore should be administered from headquarters to be most effective.

A nonsteel company with a score of 2.32 said

> In the Industrial Relations or any other field, decentralized *administration* is highly desirable. The central group can provide three important things: *Company policy manuals* to give guidance to local administrators; *staff assistance* (especially to small plants) in the development of employee safety manuals, training material, and the like which are beyond local plant staff capacity; *periodic audits* or inspections to provide assurance to top management that decentralized administration has not resulted in failure of a location to carry out important Company policies.

> We find that a high degree of central control is necessary in the negotiation and interpretation of contracts (to prevent the unions from playing one plant off against another) but that most other phases are best handled by local administration and responsibility within broad policy limits.

However, companies with somewhat similar problems, but with low activity centralization indexes, tend to deemphasize administrative consistency and to emphasize reporting, auditing, and counsel.

For example, a nonsteel company with a score of 0.94 said

> We believe the industrial relations activity should be decentralized into the individual plants to the greatest degree possible. Operations and employee relations are so inter-related that the effective administration should be exercised at the local level and essentially by the same management. The limit on the degree of decentralization depends on the effect of local actions on the other plants and divisions of the company, and the responsibilities retained by corporate management governing broad policy application and expenditures. In this area corporate industrial relations staffs function best—clearing and advising both local and corporate management.

And another with a score of 1.12 said

> A good rule of thumb: Centralize information—decentralize action.
>
> This does not mean that the Corporate Industrial Relations Department does not get into the actions which are normally decentralized. They do. But only when needed. For example, collective bargaining is decentralized, but with detailed, continuous reporting of local bargaining, the Corporate Department can and does step in (1) when the trend of the local bargaining may substantially affect more than one local unit; or (2) when unorganized units are organized for the first time; or (3) when management changes bring new, inexperienced management to the local bargaining table.
>
> This "rule of thumb" also means centralized records—but decentralized wage and salary administration—with the Corporate staff stepping in when special problems arise or major revisions of the local wage-salary program are contemplated.

In almost all the comments, the underlying theme seems to be, "Centralize policy, decentralize practice." But in working out this slogan, there seems to be a wide divergence concerning what policy embraces, which practices may be delegated, and how far.

Activity centralization across companies. An analysis was also undertaken for each of the 38 activities, based on the responsibility classifications assigned by the 42 companies that completed the scale, to identify the most frequently occurring activities and the relative degree of centralization of each activity, over these 42 companies.

For the combined sample, the five most frequently occurring corporate industrial relations functions (as indicated by the activity ratio) were: union contract negotiations, contract interpretation, supervisory training, employee benefits administration, and grievance procedure administration. Supervisory training and grievance procedure administration tend to involve advisory responsibilities

TABLE 3

ACTIVITY CENTRALIZATION INDEX AND ACTIVITY RATIO, BY ACTIVITIES (FOR 42 RESPONDING COMPANIES)

Activity	Activity centralization index	Activity ratio
Union contract negotiations	2.45	97.7
Contract interpretation	2.14	97.7
Supervisory training	1.76	92.9
Employee benefit administration	2.20	92.9
Grievance procedure administration	1.74	90.5
College recruiting	2.34	85.8
Management development	2.08	85.8
Induction and orientation training	1.51	85.8
Skills and apprentice training	1.46	85.8
Personnel records	1.95	83.4
Industrial relations research	2.63	81.0
Executive health programs	2.35	78.6
Safety training	1.55	81.0
Selection testing	1.77	78.6
Other safety activities	1.40	76.2
Occupational hygiene	1.57	73.9
University management programs	2.13	71.5
Salaried job evaluation	1.76	71.5
Employee attitude surveys	2.23	69.1
Tuition aid programs	2.22	69.1
Hourly job evaluation	1.40	69.1
House organs	1.73	66.7
First aid and medical programs	2.59	64.3
Community relations	1.05	64.3
Employee communications	2.48	62.0
Recreation programs	1.09	62.0
Test validation	1.59	59.6
Merit rating	1.39	59.6
Eating facilities	1.02	59.6
Employee counseling	1.23	57.2
Plant security	0.95	54.8
Testifying before governmental bodies	1.31	52.4
Speeches	1.29	52.4
Participation in industry associations	1.13	52.4
Public relations	1.03	47.7
Credit union	0.86	45.3
Time and motion study	0.74	31.0
Lobbying	1.24	12.0

primarily but the other three involve administrative or review responsibilities.

At the other end of the scale, the functions (among the 38 listed) that occur least frequently among corporate industrial relations staff responsibilities are: lobbying, time and motion study, credit union, and public relations. Generally, industrial relations responsibility for public relations is advisory only. This reflects the fact that the public relations function is part of the industrial relations function in only a minority of the companies surveyed. Where the two are in one department, however, the corporate staff generally exercises administrative responsibility.

From the point of view of centralization–decentralization (as measured by the activity centralization index), the five most centralized activities were: industrial relations research, first aid and medical care, employee communications, union contract negotiating, and executive health programs. The five least centralized activities were: time and motion study, credit unions, plant security, eating facilities, and community relations. For these latter activities, if the corporate industrial relations staff had a role at all, it was primarily advisory and consultative in nature.

CONCLUSION

The actual distribution of statistics of the small sample of companies surveyed is not intended to provide benchmarks, particularly for companies that are substantially smaller or substantially larger than those surveyed. It is suggested, however, that the methods and measures outlined do offer a way of reducing to common scales some very complex concepts, and collection of data over samples of companies comparable in size and structure would provide useful comparisons for a company engaged in evaluating its own industrial relations organization.

REFERENCES

AMERICAN MANAGEMENT ASSOCIATION. *Statistical analysis of the 1961 Red Book data. Group 10 research report*. Saranac Lake, N. Y.: American Management Association Academy, November 1, 1961. Unpaginated.

FRENCH, S. H., JR. Measuring progress toward industrial relations objectives. *Personnel*, 1954, 30 (5), 338–347.

GRAY, R. D. Evaluating the personnel department. *Personnel*, 1965, 42(2), 43–56.

JACKSON, W. C. The Personnel Activity Index: A new budgeting tool. *Personnel*, 1961, 38(1), 47–52.

MEE, J. F. (Ed.) *Personnel handbook*. New York: Ronald Press, 1958.

MERRIHUE, W. V., & KATZELL, R. A. ERI-yardstick of employee relations. *Harvard Business Review*, 1955, 33(6), 91–99.

NELSON, R. J., ENGLAND, G. W., & YODER, D. Personnel ratios, 1960: An analytical tool. *Personnel*, 1960, 37(6), 18–28.

NELSON, R. J., ENGLAND, G. W., & YODER, D. Personnel ratios and costs, 1962. *Personnel*, 1963, 40(1), 19–26.

WORTMAN, M. S. Personnel ratios and personnel departments. *Personnel Administration*, 1963, 26(6), 46–51.

(Received September 19, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 393–395

USE OF THE WEIGHTED APPLICATION BLANK IN SELECTING UNSKILLED EMPLOYEES[1]

RICHARD D. SCOTT AND RICHARD W. JOHNSON

University of Massachusetts

The effectiveness of the weighted application blank (WAB) in differentiating between long-term and short-term unskilled employees was evaluated. The WAB scores correlated .45 with job tenure for Ss in the holdout group (N = 50). When compared with a multiple-regression equation (r = .31 between predicted tenure and actual tenure), the WAB technique fared well. Factor analysis of the predictor variables yielded 2 factors, "convenience" and "family responsibility," which accounted for most of the explained variance in the criterion. Females who lived close to the plant and workers with a fair amount of family responsibility (e.g., married, older, several dependents, live in own home) were more likely to become long-term employees.

Although the weighted application blank (WAB) has frequently been used to hire office and sales personnel, it appears to have been used relatively little in selecting unskilled employees. Dunnette and Maetzold (1955) report one study in which the WAB was used to select unskilled workers; however, their study was concerned only with seasonal employees.

The primary purpose of this study was to develop a WAB which would be successful in selecting permanently employed workers at the unskilled level. Two related purposes were (*a*) to compare the effectiveness of the WAB with a multiple-regression equation based upon the same data in predicting employee turnover, and (*b*) to describe by means of factor analysis the personal history items associated with long-term employment.

METHOD

Subjects. A survey was made of the records of all full-time, permanently employed, unskilled workers hired by a small canning factory in western Massachusetts during the past 4 yr. The unskilled jobs consisted of such tasks as cleaning and bottling pickles. From this subject (*S*) pool, 75 long-tenure *Ss* (6 mo. or more) and 75 short-tenure *Ss* (1 mo. or less) were selected on the basis of completeness in filling in the job application form.

Biographical items. The 19 biographical items listed on the application form were used as the predictor variables.

Procedure. According to the procedures described by England (1961), two weighting groups (a long-tenure and a short-tenure group) of 50 *Ss* each were randomly selected from the larger sample of employees. Item responses for these *Ss* were compared for the purpose of assigning weights to the items. Differential weights (0, 1, or 2) were assigned as outlined by England.

The remaining 25 *Ss* in each group were placed into two holdout (cross-validation) groups to check the effectiveness of the item weights in discriminating between long- and short-term employees.

Tenure, which was of vital concern to the management, was selected as the criterion of worker effectiveness. Thousands of dollars in unfilled contracts were lost each year as the result of labor shortage due to rapid employee turnover.

The 100 *Ss* in the two weighting groups were also used to compute the multiple-regression equation between the application-blank items and the criterion. A sequence of multiple-linear-regression equations was calculated in a stepwise manner by means of the Control Data 3600 computer. The relative success of the two weighting techniques (WAB versus multiple regression) in predicting job tenure for members of the holdout groups was determined both by (*a*) percentage of correct classification, and (*b*) product-moment correlation coefficient.

The intercorrelations among the predictor items for the 100 members of the two weighting groups were factor analyzed by the method of principal components.[2] The factor matrix was graphically rotated to yield simple structure on the criterion variable. Only those factors which produced signifi-

[1] Based upon a master's thesis completed by Richard D. Scott under the supervision of Richard W. Johnson. The helpful suggestions of Stanley M. Moss and J. Alfred Southworth, thesis committee members, are gratefully acknowledged.

[2] The factor-analysis and multiple-regression programs were obtained from the series of biomedical computer programs developed at the University of California, Los Angeles (Dixon, 1965).

TABLE 1

QUARTILES, MEANS, AND STANDARD DEVIATIONS OF WAB SCORES FOR LONG- AND SHORT-TERM EMPLOYEES (HOLDOUT GROUPS)

Group	Q_1	*Mdn*	Q_3	*M*	*SD*
Short term	5	8	12.5	9.32	4.42
Long term	11	14.5	16	13.88	3.85

Note.—$N = 25$ in each group.

cant factor loadings ($p < .05$) on this variable were interpreted.

RESULTS

Of the 19 items, 12 were assigned differential weights through the use of the WAB technique.[3] When these weights were applied to members of the holdout groups, 72% correct classification of *S*s was obtained by using a cutting score of 14 (the point of maximum differentiation). This percentage of "hits" was significantly different from chance (50:50 base rate) at the .01 level of probability ($CR = 3.10$).

Table 1 presents the quartiles, means, and standard deviations for members of both holdout groups. The difference between the means was statistically significant ($t = 6.38$; $p < .01$). The product-moment correlation coefficient between the scored application blanks and months on the job was .45 ($p < .01$) for the 50 members in the holdout groups.

The multiple correlation between the 12 significant items and the criterion was .71. Nearly as high a multiple R ($R = .69$) was obtained using just six items. The multiple-regression equation based upon these six items is shown below:[4]

$Y' = .30$ (age) $+ 8.82$ (sex) $- .69$ (miles from plant) $+ 5.29$ (type of residence) $+ 2.66$ (number of children) $+ 1.08$ (years on last job) $- 1.99$.

When the application forms for members of the holdout groups were scored by means of the multiple-regression equation, the percentage of correct classification (based on a cutting score of either 9 or 10) fell slightly from 72% to 70%. The size of the correlation between the weighted scores and the criterion dropped from .45 to .31. (Neither of these differences was significantly different from chance at the .05 level.) Both indexes suggest that the WAB was as efficient as, if not more than, the multiple-regression equation in assigning weights to the variables.

Because the optimal cutting scores were established after inspection of the data (as suggested by England), accuracy of classification was also compared at several arbitary cutting levels. The relative success of the two techniques in predicting tenure status for the holdout *S*s at the twenty-fifth, fiftieth, and seventy-fifth percentile points is shown in Table 2. While the WAB appears to hold a slight edge in accuracy of prediction, the differences which exist may best be explained in terms of chance alone.

Factor analysis produced two rather-clear-cut factors which loaded high on the criterion variable: (*a*) "family responsibility" accounted for 16% of the variance in the criterion, and (*b*) "convenience" accounted for 31% of the variance. Only 4% of the variance was accounted for by the remaining factors ($h^2 = .51$); 49% of the variance was unaccounted for by any of the factors in the factor analysis.

DISCUSSION

The WAB proved to be an effective technique in selecting long-term unskilled workers. The higher the individual's score on the WAB, the greater the likelihood of his staying with the company a reasonable length of time. While the cutting score of 14 maximally differentiates between the long-term and the short-term employees, the company in practice would use as high a cutting score as the available labor market permitted.

It is surprising that weighting by the multiple-regression technique was not more effective than the WAB. Two explanations appear to be most plausible: (*a*) the use of a fairly large number of predictor variables

[3] Tables reporting the assigned weights, means, standard deviations, intercorrelations, and factor loadings for all variables may be found in R. D. Scott's (1966) master's thesis on file in the University of Massachusetts library.

[4] The discontinuous variables were coded as follows: sex: 0 = male, 1 = female; type of residence: 0 = live with parents or in room, 1 = live in own home.

TABLE 2

COMPARISON OF WAB AND MULTIPLE REGRESSION IN PREDICTING TENURE STATUS FOR HOLDOUT SUBJECTS AT 3 ARBITRARY CUTTING LEVELS

Percentile	Cutting-point score		Accuracy of classification	
	WAB	Multiple regression	WAB	Multiple regression
Twenty-fifth	7.5	5	70%	62%
Fiftieth	12	11	70	66
Seventy-fifth	15.5	16	66	66

Note.—None of the differences between the percentages at the three cutting levels was significantly greater than chance expectancy ($p > .05$). $N = 50$.

(initial $n = 19$) with a relatively small N ($N = 100$) may have yielded unstable weights, and (*b*) several of the variables (e.g., age, education, distance lived from plant, and years lived in the state) appeared to be nonlinearly related to the criterion.

Future comparisons of the two weighting procedures may profit by the use of a larger N and the application of curvilinear multiple-regression techniques. Until such comparisons can be made, however, the WAB, which is much easier to develop and apply, stands as the recommended weighting procedure.

The two factors isolated by means of the factor analysis, "convenience" and "family responsibility," appear reasonable. The one, convenience, suggests that unskilled females who live fairly close to work are likely to stay on the job. The other, family responsibility, describes the long-term employee as older, married, providing for one or more dependents, living in his own home, and as having worked a relatively long period of time on his last job.

The convenience factor agrees well with Dunnette and Maetzold's (1955) findings that females make good seasonal, unskilled employees and that, in general, workers who live near the plant are more likely to continue on the job. The family responsibility factor, however, is missing in Dunnette and Maetzold's data. Presumably, the unskilled worker who is employed only seasonally is less likely to hold large-scale family responsibilities. The long-term seasonal worker, described as young (under 25) or old (over 55), married, but with no children, and possessing 10 yr. or more education, suggests a young college student or semiretired individual in search of summer employment. This same person would be unlikely to seek unskilled employment on a permanent basis.

CONCLUSION

The WAB developed in this study successfully differentiated between long-term and short-term employees in the cross-validation study. The WAB technique appeared to be as effective as, and certainly simpler than, the multiple-regression technique in weighting the item responses.

Factor analysis of the data suggested that two dimensions, convenience and family responsibility, could adequately account for most of the variance in the criterion tapped by the variety of biographical items. The family responsibility factor appeared to differentiate between the permanently employed and the seasonally employed long-term unskilled workers.

REFERENCES

DIXON, W. J. (Ed.) *Biomedical computer programs.* (Rev. ed.) Health Sciences Computing Facility, Department of Preventive Medicine and Public Health, School of Medicine, University of California, Los Angeles, September 1965.

DUNNETTE, M. D., & MAETZOLD, J. Use of a weighted application blank in hiring seasonal employees. *Journal of Applied Psychology,* 1955, **39**, 308–310.

ENGLAND, G. W. *Development and use of weighted application blanks.* Dubuque, Iowa: William C. Brown, 1961.

SCOTT, R. D. Using the weighted application blank in determining the factors associated with job turnover. Unpublished master's thesis, University of Massachusetts, 1966.

(Received September 21, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 396–402

AN EMPIRICAL INVESTIGATION OF TWO IMPLICATIONS OF THE TWO-FACTOR THEORY OF JOB SATISFACTION

CHARLES L. HULIN AND PATRICIA A. SMITH[1]

University of Illinois

2 implications of the 2-factor theory of job satisfaction are derived and tested empirically using data from 670 office employees, supervisors, and executives employed by the same company. The results indicate that the predictions of the 2-factor theory were provided no support whatsoever. The traditional model of job satisfaction, which holds that any variable in the job situation can be both a satisfier and a dissatisfier and that if the presence of a variable tends to make a job desirable then the absence of the same variable makes a job undesirable, was supported.

One of the major propositions of the two-factor theory of job satisfaction (Herzberg, 1966; Herzberg, Mausner, & Snyderman, 1959) which distinguishes it from other, more traditional, models of job satisfaction is that certain variables in the work situation lead to overall job satisfaction but not to dissatisfaction. Other variables lead to overall job dissatisfaction but not to satisfaction. (These two classes of variables have been labeled at different times satisfiers and dissatisfiers, motivation and hygiene variables, content and context variables, and intrinsic and extrinsic variables.) In other words, satisfaction with a dissatisfier should cause no more overall satisfaction than merely being neutral with regard to the dissatisfier. Further, dissatisfaction with a satisfier should not contribute any more to overall dissatisfaction than being neutral with a satisfier.

From this proposition two implications can be drawn. If different variables contribute to job satisfaction than contribute to job dissatisfaction then Herzberg would argue that satisfaction is qualitatively different from dissatisfaction. A second implication would be that while the presence of a job characteristic will have one effect on a job the absence of the same characteristic will not have the opposite effect.

Recently two articles (Ewen, Smith, Hulin, & Locke, 1966; Graen, 1966) have presented data which indicate that, contrary to the predictions of the two-factor theory, satisfaction with a dissatisfier does lead to overall satisfaction and dissatisfaction with a satisfier does lead to overall dissatisfaction. The results of these two analyses (which were based on the same data) clearly support the traditional theory of job satisfaction without the assumption, of course, that all variables are equally potent contributors to job satisfaction. Unfortunately, both the Ewen et al. and the Graen articles have what could be considered a serious shortcoming. They measured overall job satisfaction by means of the General Motors Faces Scale (Kunin, 1955) which assumes that job satisfaction-dissatisfaction is a continuum and job dissatisfaction is merely a low level of job satisfaction. (The General Motors Faces Scale consists of drawings of faces varying from one with a deep frown through a "neutral" face to one with a broad smile. The *S* is requested to check the face which indicates how he feels about his job most of the time.) If, as the two-factor theory suggests, satisfaction and dissatisfaction are qualitatively different, the use of such a continuum could raise some doubt about the conclusions and the criticisms of Herzberg re-

[1] It should be noted that the junior author on this present article is a graduate student at the University of Illinois, Urbana. She should not be confused with Patricia Cain Smith, now at Bowling Green University, who was the senior author's major professor. This paper represents a portion of a thesis submitted to the Graduate College of the University of Illinois by the junior author in partial fulfillment of the requirements for the Master of Science degree.

sulting from these analyses. It would seem that in order to adequately test the two-factor theory it must be assumed that it is correct and satisfaction is qualitatively different from dissatisfaction, and overall satisfaction and dissatisfaction must be measured on different scales. If the Ewen et al. and Graen findings are supported after the separate measurement of satisfaction and dissatisfaction, then much more confidence could be placed in the validity of their conclusions.

In this paper the authors propose to analyze the contributions of different variables to overall satisfaction and dissatisfaction and to examine the differences between the presence and absence of different variables in their effects on workers' judgments of jobs.

METHOD

Subjects

The *Ss* for this study were the home-office personnel of a large international corporation with its home office located in Montreal, Quebec. In July 1966 all of the employees were requested to take part in a job-satisfaction survey which the company was conducting in cooperation with the authors. The questionnaires were given to the employees by departmental representatives after the representatives had met with the senior author and had the questionnaire explained to them. The employees were asked to complete them at home or, if they thought they would have 30 uninterrupted min., to do them at work. All employees up to and including the executive vice-president were asked to participate in the study. The questionnaires were then placed in envelopes, sealed by the respondents, and returned either to the investigator or sent through the mails directly to the University of Illinois. Anonymity was guaranteed to all respondents and they were assured that their individual responses would not be made known to the company. Names were not requested on the questionnaires, but for many of the employees their job title and department (which were requested) would be enough to identify the respondent. Of the approximately 800 employees who worked in the home offices and were present during the period of the study, 670 (83%) returned usable questionnaires. Not all *Ss* completed all the questions so the *N* for any given analysis may not equal 670.

Variables

Satisfaction with five job aspects (work done, pay, promotional opportunities, supervision, and co-workers) was assessed by means of the Job Description Index (JDI). The JDI is a cumulative-point adjective checklist which appears to possess high convergent and discriminant validity (Quinn & Kahn, 1967, p. 456; Vroom, 1964). All questionnaires contained the five JDI scales.

Overall job satisfaction was assessed by one of three variations of the General Motors Faces Scale (Kunin, 1955). The original scale consisted of five faces with varying smiles, a neutral face, and five faces with varying frowns. One variation of this scale utilized in this study consisted of the neutral face as its low end and the five faces exhibiting varying degrees of happiness. The workers were asked to indicate their feelings of *satisfaction* with their *job in general* (JIG) on this scale. They were asked to check the neutral or unsmiling face if they experienced no satisfaction on their job. A second scale consisted of the neutral face as the high end and the five faces with the varying frowns. On this scale the workers were asked to indicate their degree of *dissatisfaction* with their *job in general*. They were asked to check the neutral or unsmiling face if they experienced no dissatisfaction on ther job. The third scale consisted of three smiling faces, the neutral face, and two frowning faces. On this scale the workers were asked to indicate their feelings of *satisfaction* or *dissatisfaction* with their *job in general*. The numerical values assigned to the dissatisfaction scale faces were 1 through 6. For the satisfaction scale the values were 6 through 11. For the satisfaction-dissatisfaction scale the faces chosen were 2, 4, 6, 7, 9, and 11. The satisfaction and the dissatisfaction scales overlap at one point, that is, the neutral face which was given a value of six. The satisfaction-dissatisfaction scale overlapped with both the satisfaction and the dissatisfaction scales. Any face was given the same numerical value no matter which scale it appeared in. There is evidence available (Kunin, 1955) that the numerical values assigned to these faces approximate an equal interval scale. It should be noted that the satisfaction-dissatisfaction scale used in the present study is identical to the scale used in the Ewen et al. and Graen studies and thus provides a partial replication of their results. One-half of the workers received the satisfaction-dissatisfaction scales, one-quarter the satisfaction scale, and one-quarter the dissatisfaction scale. The scales were randomly distributed within departments.

Two other scales were designed expressly for this study to assess the contribution of the presence or absence of various job characteristics to the manner in which workers responded to different jobs. Six different working situations were described in terms of six job characteristics (interest and difficulty of work done, working conditions, pay received, promotional opportunities, co-workers, and supervision). On one of these two scales, each working situation would be described as having five of the characteristics about average (e.g., the work is about average in interest and difficulty, the working conditions are about average in comfort and convenience, etc.), and one of the characteristics would be described as being outstandingly good (e.g., your co-workers are

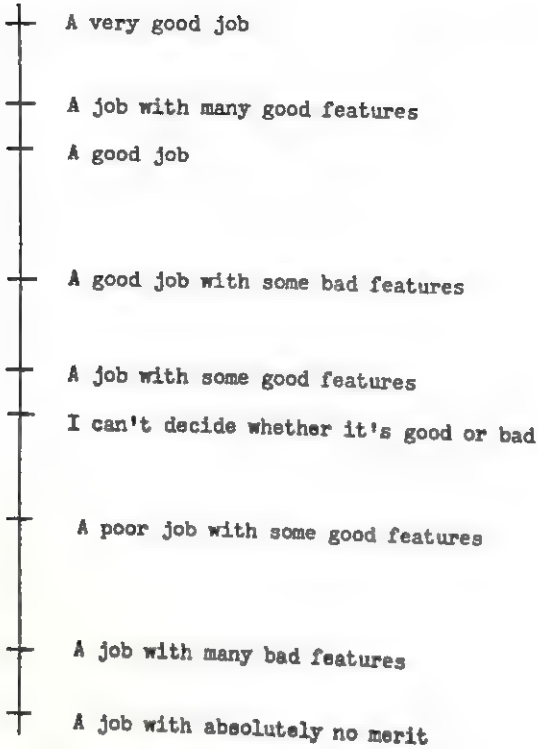


FIG. 1. Graphic rating scale for obtaining judgments of desirability of various working situations.

very enjoyable to work with). Each characteristic appeared once as being outstandingly good and five times as being average in the six working situations which were presented. The order in which the characteristics were presented in any description was randomly determined as was the order in which they were the outstanding characteristic. The *Ss* were asked to indicate on an anchored graphic rating scale shown in Figure 1 how they felt about work on such a job, how they would describe it to their friends. Scale values for each of the adjective phrases were calculated by the method of categorical judgments, Condition B, using a graphical solution (Torgerson, 1958) on an independent sample of clerical workers employed by the University of Illinois. The assumptions of this scale were that the working situations described by the respondents as being good or desirable would tend to be those which possessed an outstanding amount of an important characteristic. Thus, the determination of those *situations* which were consistently judged as being good or desirable would enable one to determine the relative importance of those *job characteristics* whose presence in outstanding amounts has an impact on the way workers react to jobs.

A second version of this scale consisted of similar descriptions of the working situations except in this latter version five of the characteristics were average and one was bad (e.g., you have almost no chance to be promoted). Again the workers were asked to indicate on the same scale how they would feel about such a job. By reasoning similar to that above, this version of the scale should enable us to determine those job characteristics whose absence would have an impact on workers' reactions to their jobs. Approximately one-half of the workers received each version of this scale. The forms were randomly distributed within departments.

Predictions

Based on these 12 variables (five JDI satisfaction variables, JIG, and the six job aspects) a number of specific hypotheses can be generated where the traditional model of satisfaction and the two-factor theory make antithetical predictions. The two-factor theory would predict that satisfaction with work done and satisfaction with promotional opportunities should be related to satisfaction with the job in general but not related to dissatisfaction with the job in general. The traditional model of job satisfaction would predict that satisfaction with work done and with promotional opportunities would be related to both satisfaction and dissatisfaction with the job in general. On the other hand, pay satisfaction should be related only to overall dissatisfaction and not to overall satisfaction according to the two-factor theory. The traditional model would predict that pay satisfaction would be related to both satisfaction and dissatisfaction. The other two areas of satisfaction measured by the JDI do not permit clear predictions. Supervision is supposedly a dissatisfier, but supervision may contain elements of recognition which is a satisfier. Satisfaction with co-workers at times has appeared to be a dissatisfier (Herzberg, 1966, p. 97, 102, 108), but at other times has operated as a satisfier (Herzberg, 1966, p. 104). Therefore no predictions were made for these two variables based on the two-factor theory. The traditional model would predict that both of these variables would be related to both satisfaction and dissatisfaction.

The two-factor theory would make no predictions regarding the overall potency of the satisfaction areas being measured in this study. However, based on the results of the Ewen et al. and Graen studies, it could be predicted that satisfaction with work done and promotional opportunities would be more highly related to overall satisfaction than would pay satisfaction. Since these were the only three job aspects dealt with in the two studies no predictions can be made regarding the relative potency of all five of the aspects.

In terms of the impact of the six job characteristics used to describe the six working situations on the way workers report they would feel about a job, clearly antithetical predictions can again be generated. The traditional model would predict that if the presence of a variable was related to a job being described in good terms (being judged as a good job), then if that variable were absent the job would be described in relatively poor terms (being judged as a bad job).

Thus, if the importance or impact of the six job characteristics as computed from the form of the questionnaire where five characteristics were average but one was outstandingly good were plotted against the values computed from the form where five characteristics were average and one was outstandingly bad, a positive linear function (within the limits of the equality of the scaling intervals) should be obtained. The two-factor theory, on the other hand, would make no such prediction. Since the presence of certain characteristics leads to a desirable job but the absence of the same characteristics does not lead to the job being considered a bad job one would not expect such a function. The two-factor theory would predict that if the work were very interesting or if there were good opportunities for advancement on jobs, then these jobs would be seen as very good, other things being equal. However, if the work were very uninteresting or if there were no opportunities for advancement these jobs would not necessarily be judged as being bad jobs, other things being equal. They would only be neutral on the scale. Low pay, on the other hand, should lead to a job being judged as being relatively bad, but high pay should not necessarily lead to a job being judged as being good. Thus, the two-factor theory would not predict a linear positive function between the importance of the six characteristics as obtained from the two forms of the questionnaire. The form of the function would be indeterminant from the assumption of the two-factor theory.

RESULTS

The results of the correlational analysis for the male employees are presented in Table 1. Looking only at the three variables for which predictions were made evidence is found clearly supporting the predictions of the traditional model of job satisfaction and clearly disconfirming the predictions of the two-factor model of job satisfaction. Satisfaction with work done, pay received, and promotional (advancement) opportunities and policies is significantly related to satisfaction, dissatisfaction, and satisfaction-dissatisfaction with the job in general (JIG). The lower correlations between these variables and JIG dissatisfaction are probably due to the sharply reduced variance on this latter scale. Due to the satisfaction levels encountered in this company the JIG dissatisfaction scale was essentially a 3-point scale (*Ss* checked mainly the three top faces with only two *Ss* checking Face No. 3). Nonetheless the correlations between work and promotion satisfaction (supposedly satisfiers) and JIG dissatisfaction were significant as was the correlation between pay satisfaction (essentially a dissatisfier) and JIG satisfaction. These three highly significant correlations are directly opposed to the predictions of the two-factor model.

The correlations between supervision and co-worker satisfaction and the JIG scales are presented in the interest of hypothesis formulation since no predictions were made for these two variables.

The correlations between work, pay, and promotion satisfaction and JIG satisfaction-dissatisfaction lend but slight support to the

TABLE 1

CORRELATIONS BETWEEN SATISFACTION WITH DIFFERENT ASPECTS OF THE JOB AND OVERALL JOB SATISFACTION FOR MALE EMPLOYEES

JDI satisfaction area	JIG satisfaction-dissatisfaction ($N = 170$)	JIG satisfaction ($N = 95$)	JIG dissatisfaction ($N = 84$)
Work	.51**	.68**	.44**
Pay	.32**	.39**	.24*
Promotion	.42**	.40**	.38**
Supervision	.32**	.53**	.25*
Co-workers	.37**	.48**	.13

* $p < .05$.
** $p < .01$.

TABLE 2

CORRELATIONS BETWEEN SATISFACTION WITH DIFFERENT ASPECTS OF THE JOB AND SATISFACTION WITH JOB IN GENERAL FOR FEMALE EMPLOYEES

JDI satisfaction area	JIG satisfaction-dissatisfaction ($N = 154$)	JIG satisfaction ($N = 52$)	JIG dissatisfaction ($N = 68$)
Work	.60**	.45**	.43**
Pay	.30**	.12	.18
Promotion	.41**	.46**	.14
Supervision	.45**	.31*	−.03
Co-workers	.36**	.20	−.08

* $p < .05$.
** $p < .01$.

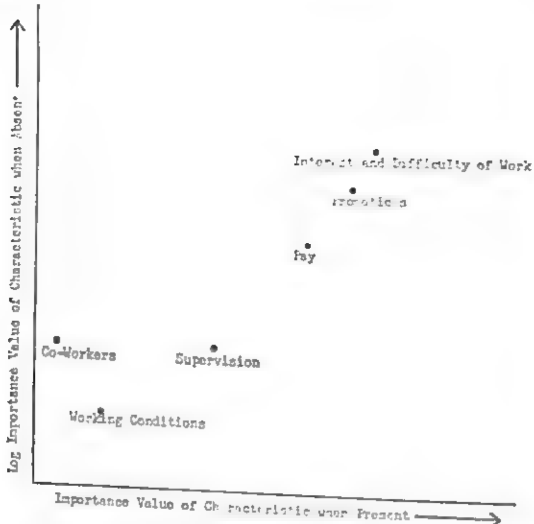


FIG. 2. Importance values of job characteristics when present versus importance values of job characteristics when absent, male sample.

generality of the previous results of Ewen et al. and Graen regarding the relative potency of the intrinsic as opposed to the extrinsic variables. Both work and promotion satisfaction are more highly correlated with JIG satisfaction-dissatisfaction than is pay satisfaction. However, only the work satisfaction/JIG satisfaction-dissatisfaction correlation is significantly higher than the pay satisfaction/JIG satisfaction-dissatisfaction correlation.

The results of the correlational analysis based on data from the female *Ss* are presented in Table 2. As is typical, these data from the female *Ss* are less orderly than those from the male sample (cf. Hulin & Smith, 1964, 1965). These results, however, do not support the predictions of the two-factor model since the only variable which is significantly correlated with JIG dissatisfaction is work satisfaction (supposedly a satisfier). Pay satisfaction, contrary to the predictions of the two-factor theory, is unrelated to JIG dissatisfaction. Promotion satisfaction, on the other hand, does fit the predictions of the two-factor theory. Again it should be pointed out that the variance on the JIG dissatisfaction scale was much lower than the variance of either of the other two JIG scales. This difference could result in attenuated correlations. However, the pattern of correlations which was obtained from the female sample lends little support to the two-factor theory. Again, work satisfaction is more highly correlated with JIG satisfaction-dissatisfaction than is pay satisfaction. The promotion satisfaction/JIG satisfaction-dissatisfaction correlation is not significantly greater than the corresponding value for pay satisfaction.

Neither of these two sets of data supports the predictions of the two-factor theory. The data from the male sample do support the predictions of the traditional model and the female data partially support these predictions.

The results of the importance analysis from the two forms of the questionnaire from the male sample are given in Figure 2. In both Figure 2 and Figure 3 the log of the importance or impact value of the job characteristic if absent from the working situation is plotted against the importance or impact value of the job characteristic if present in the working situation. Using this log transformation on the scale values of the importance of the characteristic if absent merely corrects for unequal scale intevals at the two ends of the scale shown in Figure 1. Values of characteristics when absent are based on a sample of 180. Values of characteristics when present are based on a sample of 191. The relationship between the two sets of values in Figure 2 is obviously linear (under the transformation described above) and the fit is good. Thus, these results clearly indicate that, for male workers, if the presence of a variable results in a job being described or judged as good, the absence of that same variable results in the job being described as bad, other things being equal. This finding, of course, supports the traditional model of job satisfaction and argues against the two-factor theory. Only the values for satisfaction with co-workers are seriously out of line with the prediction of the traditional model.

The results of the importance analysis of the female sample are given in Figure 3. Importance values of characteristics if present are based on a sample of 157. Importance values of characteristics if absent are based on a sample of 140. As in the correlational

analysis these data from the female *S*s are less clear-cut than the male data. While a regression line through these data points would be positive and linear, there is more scatter about the line than in the male data. It is interesting to note that interest and difficulty of the work, pay level, and promotional opportunities have nearly the same effects if present or absent as they did in the male sample. For the other three variables, the results are more random. Again, these results do not support the two-factor theory and tend to confirm the predictions of the traditional model. Also, satisfaction with co-workers does not appear to fit well with the remainder of the variables.

The data in Figures 2 and 3 also support the findings of Ewen et al. and Graen in that they clearly indicate differential saliency among the various factors which contribute to job satisfaction. Further, in both the male and female samples the intrinsic aspects of the job (interest and difficulty of the work) are responded to by the workers as being most important. This also supports the conclusions of Ewen et al., in that a division of the factors into intrinsic and extrinsic is more meaningful than the labels "satisfiers" and "dissatisfiers."

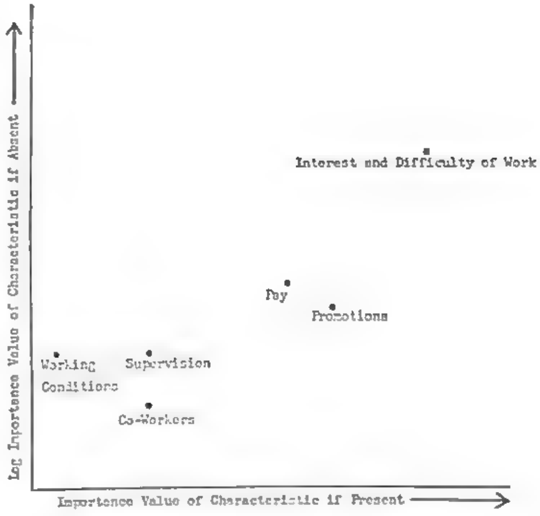


FIG. 3. Importance values of job characteristics when present versus importance values of job characteristics when absent, female sample.

DISCUSSION

It is clear that the present results provide no support for the predictions which one would make on the basis of the two-factor theory of job satisfaction. The so-called satisfiers acted as both satisfiers and dissatisfiers, and the dissatisfier acted as satisfier as well as a dissatisfier. The analysis of the importance of various job characteristics on the way people reacted to different job situations was the same whether the characteristic was present or absent. No evidence was found which would support the argument that satisfaction and dissatisfaction are qualitatively different. If the differences in the variances between the JIG satisfaction and the JIG dissatisfaction scales are taken into account, these scales do not give qualitatively different results than does the JIG satisfaction-dissatisfaction scale in terms of their patterns of correlations with other variables. It would seem that the traditional model of job satisfaction has been strongly supported at the expense of the two-factor theory.

With these results, along with the other failures to replicate Herzberg's original results whenever different research methods are used (see Dunnette, Campbell, & Hakel, 1967; Ewen, 1964; Ewen et al., 1966; for a discussion of these nonconfirmatory results and the methods used), it would be reasonable to point out again that Herzberg's (1959, 1966) results appear to be method bound and the conclusions appear to pivot on method variance rather than true content or scale variance. A construct which can be generated or supported by only one operation would seem to have little relevance to the behavior of workers. Little more can be said. The authors join with Dunnette, Campbell, and Hakel in hoping that the two-factor theory can be laid to rest with a minimum of fanfare.

REFERENCES

DUNNETTE, M. D., CAMPBELL, J. P., & HAKEL, M. D. Factors contributing to job satisfaction and dissatisfaction in six occupational groups. *Organizational Behavior and Human Performance*, 1967, 2, 143–174.

EWEN, R. B. Some determinants of job satisfaction: A study of the generality of Herzberg's theory. *Journal of Applied Psychology*, 1964, 48, 161–163.

EWEN, R. B., SMITH, P. C., HULIN, C. L., & LOCKE, E. A. An empirical test of the Herzberg two-factor theory. *Journal of Applied Psychology*, 1966, **50**, 544–550.

GRAEN, G. B. Addendum to "An empirical test of the Herzberg two-factor theory." *Journal of Applied Psychology*, 1966, **50**, 551–555.

HERZBERG, F. *Work and the nature of man*. New York: World, 1966.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. *The motivation to work*. New York: Wiley, 1959.

HULIN, C. L., & SMITH, P. C. Sex differences in job satisfaction. *Journal of Applied Psychology*, 1964, **48**, 88–92.

HULIN, C. L., & SMITH, P. C. A linear model of job satisfaction. *Journal of Applied Psychology*, 1965, **49**, 209–216.

KUNIN, T. The construction of a new type of attitude measure. *Personnel Psychology*, 1955, **8**, 65–78.

QUINN, R. P., & KAHN, R. L. Organizational psychology. *Annual Review of Psychology*, 1967, **18**, 437–466.

TORGERSON, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

VROOM, V. *Work and motivation*. New York: Wiley, 1964.

(Received October 5, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 403–410

EFFECTS OF INEQUITY PRODUCED BY UNDERPAYMENT ON WORK OUTPUT, WORK QUALITY, AND ATTITUDES TOWARD THE WORK[1]

EDWARD E. LAWLER, III AND PAUL W. O'GARA

Department of Administrative Sciences, Yale University

This study provides a further test of Adams' theory of equity. Ss on a piece-rate plan were paid with an equitable wage (25¢) or an unfairly low wage (10¢) for doing an interviewing job. As predicted, the results showed that the underpaid Ss produced more interviews than the equitably paid Ss, but the interviews were of lower quality. The underpaid Ss tended to see the job as more interesting than did the equitably paid Ss, but they saw it as less important and challenging. Each S was given the California Personality Inventory (CPI), and the relationships among the CPI scale scores and Ss' job attitudes and job performance were considered. The results of the study generally supported equity theory but suggested that further elaboration is needed if it is to predict what inequity-reduction methods will be chosen by a given individual.

The focus of the present study is upon the human tendency to seek cognitive consistency. Specifically, it attempts to further test Adams' (1963a) theory of inequity. Adams' theory can be thought of as a special case of Festinger's (1957) cognitive dissonance theory. Although Adams' theory was designed to be a general theory of social inequity using the exchange model (Blau, 1964; Homans, 1961), so far the focus has been upon experimentally testing it with respect to the effects of wage inequity.

Adams (1965) defines inequity as follows:

> Inequity exists for Person whenever he perceives that the ratio of his outcomes to inputs and the ratio of Other's outcomes to Other's inputs are unequal, either (a) when he and Other are in a direct exchange or (b) when both are in an exchange relationship with a third party and Person compares himself to Other [p. 280].

The theory further stipulates that the greater the inequity of the input-outcome balance, the greater will be the resulting dissonance and the greater will be the motivation for corrective acts. Inputs in the job situation for a person may include his education, skill, and how hard he works. Outcomes include such things as pay, fringe benefits, status, and the intrinsic interest of the job. It must be remembered that what determines the equity of a particular input-outcome balance is the individual's perception of what he is giving and receiving, and this cognition may or may not correspond to the perception held by others.

Adams and his associates in testing equity theory have focused upon how inequity due to overpayment affects the quality and quantity of work in a job situation. The results of these studies have consistently tended to support equity theory; when subjects (*S*s) have been made to feel overpaid they appear to have adjusted their work quality and quantity in such a way that an equitable balance was created between their inputs and their outcomes. For example, hourly paid *S*s who were given the impression by the experimenter (*E*) that they were unqualified for the job (low inputs) worked harder (higher quantity) than did those *S*s who were told that they were qualified for the job (Adams, 1963b; Adams & Rosenbaum, 1962). This finding supports equity theory since working harder is one way of increasing inputs in order to create a favorable balance that was not present initially.

[1] The authors would like to thank C. Argyris, K. Weick, and V. H. Vroom for their helpful comments on an earlier version of this paper. T. Young and C. Koplin aided the authors by serving as experimenters for some of the subjects. Financial support was provided by an International Business Machines Corporation grant to the Department of Industrial Administration, Yale University.

In other studies (Adams, 1963b; Andrews, 1965) *Ss* have been made to feel either equitably paid or overpaid on a piece-rate basis and a quite different result appeared. The *Ss* who felt overpaid produced less than those who felt equitably paid, but the work was of higher quality. This finding is also congruent with the theory since, by restricting their productivity and increasing their work quality, the overpaid *Ss* were behaving in such a way as to lower their financial outcomes, thereby bringing them more in line with their inputs. Other studies (Adams & Jacobsen, 1964; Arrowood, 1961) have attempted to establish that the same results appear even where job security does not concern *Ss*.

Thus, the evidence generally supports the view that, when faced with overpayment, *Ss* react by altering the quality and quantity of their work in order to bring about an equitable input-outcome balance. However, the study of inequity due to underpayment has been virtually ignored. The most interesting application of equity theory here is to the situation of underpayment on a piece rate. Presumably, underpayment on a piece rate would result in large amounts of low-quality work, since this would raise the person's outcomes without increasing such inputs as effort expended. Andrews (1965) found that underpaid *Ss* when compared with equitably paid *Ss* produced somewhat more, but that there was not a significant work-quality difference. However, this may have been due to the relatively small pay difference (15¢ versus 20¢) that existed between the two groups. Thus one objective of the present study is to increase the magnitude of underpayment in order to create a more definitive test of equity theory.

It should be noted that in his work Adams has always induced feelings of inequity by the instructions he has given to *Ss*, making them feel either qualified or unqualified for the job. The Andrews study on the other hand produced the feelings of inequity by changing the actual pay rates of *Ss*. With the Adams approach the feelings of inequity are created by *E* before *S* begins the job, and the question is, how will he behave in order to reduce these feelings? When the amount of piece-rate pay is altered as was done by Andrews and as is also done in the present study, *S* does not know that he is being unfairly paid until he actually begins to perform the job. At this point the key issue concerns how the *S* will define the job in order to deal with the inequity that has been built into the situation in which he finds himself. One obvious advantage of this approach is that it provides a better simulation of the situation in which most employees who are on piece-rate plans actually find themselves. In addition, it eliminates any questions concerning the possibility that the results can be explained in terms of job insecurity, which exists where workers are told they are unqualified for the job. On the negative side, manipulating actual pay in one sense is not so clean as manipulating qualifications because *E* has less control over when and how *S* comes to realize that he is in an inequitable situation. At this point it would appear to be logical to continue collecting data using both approaches to inducing feelings of inequity, since both can provide meaningful data relevant to the theory.

Adams has emphasized that an important part of equity theory is that there are other ways of reducing dissonance due to inequity than simply varying productivity and quality of job performance. Inequity can be reduced by altering or perceptually distorting inputs and outcomes other than pay and productivity, by leaving the field, or by changing to another comparison object. Except for the study by Andrews (1965), no attention has been paid to methods of dissonance reduction that are less obvious than simply changing productivity and work quality on the input side and pay on the outcome side. Andrews found that underpaid *Ss* on a piece rate did show a significant tendency to see the job as simpler than did the equitably paid *Ss*. Presumably, seeing the job as simple made the low pay more tolerable by reducing *S*'s job inputs on a dimension (job complexity) relevant for the determination of pay. Because of the importance of considering alternative dissonance-reduction methods, it was decided in the present study to include measures of several kinds of attitudes and perceptions of job inputs and outcomes that might serve as dissonance-reduction mechanisms for

the underpaid *Ss*. The expectation was that by seeing the job as relatively simple, unchallenging, and unimportant, *Ss* that are underpaid can lower their perceptions of their inputs in a way that will reduce the inequity they feel from being underpaid. In effect, they can reason that unimportant jobs do not deserve high pay. In addition, it was expected that by seeing the job as interesting, underpaid *Ss* can increase their outcomes, thereby reducing the dissonance produced by low-pay outcomes. It may appear that it is unreasonable to predict that an *S* will say a task is at the same time simple and unchallenging, but interesting. However, the prediction is only that the underpaid *Ss* will see the task as relatively less challenging and complex, but more interesting, than will the equitably paid *Ss*. The theory does not necessarily predict that the underpaid *Ss* in an absolute sense will say the job is unchallenging and simple, but interesting.

Completely unexplored in the area of equity theory is the question of what factors influence the choice of methods used to reduce inequity. Equity theory makes few predictions here and none of the experimental work has dealt with this topic. At least one dimension that seems to be relevant is the importance to the *S* of the pay which he is earning. Presumably an *S* in a piece-rate underpayment situation who badly needs money at the time of the study will be more likely to use the dissonance-reduction method of raising his productivity and thereby his pay, than will be the *S* to whom money is relatively unimportant.

In addition to the importance of pay, there may be certain personality dimensions that influence the kind of dissonance-reduction method an individual will use in the underpayment situation. Thus, it was decided to include in the present study a personality test in order to help understand the choices of dissonance-reduction mechanisms made by *Ss* in the underpayment situation. Particularly relevant for this purpose seem to be the scales from the California Personality Inventory (CPI) that are designed to measure poise, ascendancy, and self-assurance, and those designed to measure maturity and responsibility. The self-assurance scales upon which low scorers are characterized as being self-defensive and given to feelings of guilt and self-blame appear to be relevant because as Weick (1965) has pointed out, low rewards can be quite threatening to a person's self-confidence. Because of this it seems that high-self-assurance persons might react differently to an underpayment situation than low-self-assurance persons. The responsibility scales appear to be relevant because it was expected that they might be related to the degree to which *Ss* are willing to sacrifice quality for quantity.

METHOD

Subjects

Forty Yale University undergraduate students served as *Ss* for the study. They were hired either through the student placement service or through an advertisement placed in the student post office. All *Ss* prior to applying for the experiment were told that they were being hired to work on an interviewing project and that they were being hired for only 2 hr. of work. They were also informed that their salary would be based upon a piece-rate system. The *Ss* were randomly divided into an equitable payment group ($n = 20$) and an underpayment group ($n = 20$). A comparison of these two groups on such factors as age, year in school, and expected wages showed no significant differences.

Procedure

A pretest was run to determine the size of the piece rate to be paid the equitable and the underpaid groups. The pretest *Ss* were paid on an hourly rate ($1.65) and were instructed to do interviews for 2 hr. using the same interview forms that were used for the experimental *Ss*; the pretest *Ss* averaged 15 interviews in the 2-hr. period. On the basis of this finding and the fact that the standard student placement office minimum rate was $1.65/hr for the students hired, it was felt that a rate of $.10/interview would be perceived as underpayment and a rate of $.25/interview would be perceived as equitable payment. These perceptions of equity and inequity were expected to result when *Ss* compared their input-outcome ratio with that received by other students doing part-time work at the University. Confirmation of this expectation was obtained in a questionnaire completed by each *S* upon completion of the job. The *Ss* were asked to answer an item on a 7-point scale which had as one end of the scale underpaid and at the other end, equitably paid. The underpaid *Ss* described their jobs as relatively underpaid (mean score 3.2), while the equitably paid *Ss* described their jobs as being equitably paid (mean score 4.1). The fact that differences appeared even after the underpaid *Ss* had an opportunity to reduce their feelings of inequity argues strongly that feelings of inequity were successfully induced in the underpayment group.

TABLE 1

PRODUCTIVITY, WORK QUALITY, AND ATTITUDE SCORES OF THE EQUITABLY AND THE UNDERPAID SUBJECTS

	Underpaid ($n = 20$) (10¢)	Equitably paid ($n = 20$) (25¢)	t
Productivity (No. interviews)	29.6	20.0	4.42***
Quality (No. words per interview)	42.9	60.6	2.91***
Qualifications (1 low to 7 high)	5.4	5.8	1.02
Boring–Interesting (1–7)	6.1	5.1	3.06***
Unimportant–Important (1–7)	3.6	4.4	1.86*
Simple–Complex (1–7)	3.4	4.1	1.52
Unchallenging–Challenging (1–7)	3.6	4.5	2.07**

* $p < .10$.
** $p < .05$.
*** $p < .01$.

When the experimental Ss arrived at a building located away from the psychology department to participate in the study, every effort was made to simulate an actual job situation. The Ss were first asked to complete a job application form that included general preemployment information, such as age, semester in school, and previous work experience, and also asked them to indicate their expected wage level.

Upon completion of the job application, each S was given the following job description:

> You will be paid for about three hours of your time, two hours of which we would like to have you use in interviewing fellow students about controversial topics concerning Yale. Since this is a new job and we are attempting to structure it as best possible, we would appreciate it if you would answer a few questions about the job when you have finished the interviewing.
>
> We are sorry that we cannot explain to you the precise purpose of the project, but as you must realize, this might bias your data collection technique. Since the interview consists of four open-ended questions, we are employing as many interviewers as possible in order to avoid any bias that any one interviewer might project. You will work for two hours and will be paid 25¢/10¢ for each interview.

The four questions that composed the interview were open-ended ones concerned with student social life at Yale. The questions dealt with topics such as drinking at Yale and visiting hours for women in students' rooms. Each question was designed so that the interviewer would have considerable latitude in how completely he recorded each respondent's answer. The interviewers' guide gave several general guides to effective interviewing (e.g., ask questions in order shown on form and interview only registered students). The instructions also told Ss to "record each interviewee's comments in sufficient detail to enable someone else to interpret correctly what was said." Because of the stress on recording an easily interpretable record of what was said, it was felt that the number of words recorded by the interviewer from each interview would serve as a reasonable measure of the quality of the interview. A content analysis of a random sample of interviews confirmed this expectation. More words did appear to be clearly associated with conveying more information about what took place in the interview. Adams (1963b) and Andrews (1965) have also used this measure of quality.

When Ss returned after 2 hr. of interviewing, they were asked to complete a postwork questionnaire because "this is a new job and we are considering how it might be changed for future interviewers." The postwork questionnaire asked them to indicate on two scales running from 1 (not at all) to 7 (very much) how well qualified they were for the job, and how much they needed the money from the job. In addition, the questionnaire asked them to answer five 7-point semantic differential items concerning the job (boring–interesting, unimportant–important, simple–complex, unchallenging–challenging, underpaid–overpaid).

Upon completion of the postwork questionnaire, each S was informed about the nature of the project. He was then asked not to discuss its contents with anyone, and Es stressed the point that the value of the study was dependent upon the extent to which future Ss felt they were in an employment situation. The S was then told that he would receive a payment of $5.00 or what he had earned, whichever was greater. The S was then asked if he would assist E by completing a personality test that was "designed to help us understand peoples' reactions to job situations." All Ss then completed the California Personality Inventory (Gough, 1957).

RESULTS

It was expected from equity theory that the underpaid Ss would tend to reduce the dissonance associated with underpayment by producing more than the equitably paid group did and by producing work that was of lower quality than that produced by the equitably paid group. Producing more in a piece-rate situation has the obvious consequence of raising Ss' financial outcomes, while lowering quality tends to reduce their inputs. Table 1 shows that the data from the present study are congruent with this expectation. There was a significant tendency for the underpaid

TABLE 2

PEARSON PRODUCT-MOMENT CORRELATIONS BETWEEN THE CALIFORNIA PERSONALITY INVENTORY SCALES AND THE SCORES ON JOB PERFORMANCE AND JOB ATTITUDES

	Poise, Ascendancy, and Self-Assurance					
	Dominance	Capacity for Status	Sociability	Social Presence	Self-Acceptance	Well-Being
Productivity	−.06	−.24	−.36*	−.49**	−.25	−.39*
Quality	.02	.25	.34	.27	.15	.26
Boring–Interesting	−.07	−.10	−.19	.02	−.09	.04
Job-input attitudes[a]	−.18	−.09	.03	.02	.26	.11

	Socialization, Maturity, and Responsibility					
	Responsibility	Socialization	Self-Control	Tolerance	Good Impression	Communality
Productivity	.24	.12	−.05	−.32	−.54**	.19
Quality	.38*	.18	.16	.60**	.54**	.18
Boring–Interesting	−.28	−.35	.10	.07	.30	−.33
Job-input attitudes[a]	.18	−.10	.10	.30	.26	−.10

[a] Average correlation coefficients for importance, complexity, and challenge scales.
* $p < .05$, one-tailed test.
** $p < .01$, one-tailed test.

Ss to do more interviews than the equitably paid Ss, but to record significantly fewer words for each interview completed.

Table 1 also shows that there are some significant attitude differences between the equitably paid and the underpaid groups. According to equity theory, adjustments in attitudes toward job inputs and job outcomes offer a potential method of dissonance reduction. Specifically, it was expected that dissonance might be reduced by the underpaid Ss feeling that they received significant intrinsic task-interest outcomes from the task. This expectation is supported by the data since the underpaid Ss did tend to see the job as significantly more interesting than did the equitably paid Ss. A second expectation was that the underpaid Ss might reduce their dissonance by perceiving that they had relatively low inputs to the job. This expectation was also generally supported by the data. They saw their own qualifications as less than did those Ss who were in equitably paid groups, although this difference was not significant. They also saw the task as demanding lower inputs on their part. In comparison to the equitably paid group, they saw the task as relatively unimportant, simple, and unchallenging. In summary, it appears that underpaid Ss tended to reduce their dissonance by increasing their productivity, by seeing the job as interesting, and by seeing the job as being relatively unimportant and unchallenging.

The data relevant to the question of what personality factors contribute to the degree to which different dissonance-reduction mechanisms are used by the underpaid Ss can be found in Table 2. The scales from the CPI that were designed to measure poise, ascendancy, and self-assurance show a consistent tendency to be related to productivity (three of the correlations are statistically significant and all six are in the same direction). The Ss high on these scales are lower producers than Ss who score low on these scales. No consistent relationship appears between the poise, ascendancy, and self-assurance scales and any of the attitude measures. The group of scales on the CPI that were designed to measure socialization, maturity, and responsibility show a consistent tendency to be positively related to work quality (three of the six correlations are statistically significant

and all six are in the same direction). However, they fail to show a consistent relationship to any of the other measures. Not shown in Table 2 is the correlation between the degree to which *S*s said they needed the money and their productivity; as expected, it proved to be positive and statistically significant ($r = .46$, $p < .05$). In summary, it appears that the poise and self-assurance scales, along with the importance of pay to the *S*, predicted productivity well, while the maturity and responsibility scales predicted work quality.

DISCUSSION

The results of the present study consistently tend to support the predictions of equity theory about the effects of underpayment on *S*'s job performance and job attitudes. Specifically, the results indicate that when faced with an underpayment piece-rate situation, *S*s behave in a way that is designed to increase their outcomes while it decreases their inputs. This was accomplished in the present study not only by *S*s increasing their productivity while reducing their work quality in order to increase their outcomes, but also by the kinds of attitudes and perceptions they developed toward the job.

Apparently, as soon as the underpaid *S*s realized how low their pay was, they adjusted their perceptions of what was expected of them in terms of quality and productivity in order to create a situation in which they could obtain as favorable an input-outcome balance as possible. Some secondary data were collected to support this interpretation, since each *S* was asked to record how many interviews he completed in the first hour of his work. The underpaid *S*s showed a consistent tendency to increase their output from the first to the second hour of their work, while the equitably paid *S*s showed little change from the first to the second hour of work. Although not too much weight should be put upon these data because it was impossible for *E* actually to measure productivity, they do give some idea of what changes occurred after *S*s realized the kind of payment situation in which they found themselves.

It is interesting to note that with a mean production rate of 29.6 interviews, the underpaid *S*s earned an average of \$1.48/hr. This is slightly below the expected wages for placement office jobs at Yale and, as indicated, these *S*s did feel underpaid. The equitably paid *S*s, on the other hand, earned \$2.50/hr which exceeded the standard placement office rate, and yet they indicated that they were equitably paid. That this group felt equitably paid at a rate of \$2.50 is not surprising, however, since Andrews (1965) has shown that the threshold for overpayment appears to be quite high, so that employees are relatively insensitive to overcompensation. Adams (1963b), for example, has used an hourly rate of \$3.50 as an equitable rate for college students doing interviewing tasks. Just the opposite condition appears to exist with undercompensation, since employees appear to be very sensitive to any condition that might involve underpayment. Also of interest is the obviously larger productivity that appeared in both the equitably paid and the underpaid groups in comparison to the pretest *S*s who averaged only 15 interviews in a 2-hr. period. The obvious difference between the groups was that the pretest *S*s were on an hourly rate, while the experimental *S*s were on a piece rate. Apparently piece rates can produce higher productivity in some situations, a finding that does not appear to be easily handled by equity theory.

The underpaid *S*s tended to feel the job was more interesting, thereby increasing their outcomes on the intrinsic task-interest dimension. The underpaid *S*s also tended to see the job as relatively unimportant, simple, and unchallenging. Seeing the job as low on these dimensions would appear to reduce *S*s' inputs since clearly a simple, unimportant, and unchallenging task demands less of an *S* than does a complex, challenging, and important one.

Weick (1964) has found a task-enhancement effect that appears to be similar to the tendency in the present study for the underpaid *S*s to describe the task as more interesting than did the equitably paid *S*s. Weick found that *S*s who find themselves in a low-reward situation tend to see the job as more interesting than do *S*s who find themselves in an equitable-reward situation. The major difference between the two studies is that the

Weick study did not offer *S*s a chance to gain greater financial outcomes by increasing their productivity as was possible in the present study. The fact that a task-enhancement effect appeared in the present study where alternative means of dissonance reduction were easily available would seem to argue that it is a powerful and rather frequently used dissonance-reduction mechanism, especially if the task is rich in content, which the interviewing task is.

It is interesting to note that the underpaid *S*s recorded slightly more words during their 2 hr. of interviewing than did the equitably paid *S*s. This would seem to indicate that the underpaid *S*s worked as hard if not harder than did the equitably paid *S*s. Weick (1964) also found that underrewarded *S*s worked as hard as equitably rewarded ones and in fact he found a tendency for them to work harder. This finding does not appear to be easily predictable from equity theory. The important difference between the two payment groups in the present study was in how many words the *S*s recorded for each interviewee. Clearly the underpaid *S*s were willing to sacrifice quality by getting few words per interview in order to increase their monetary outcomes.

Thus, when considered together, the results of the comparison between the reactions to the job by the equitably paid *S*s and the underpaid *S*s offer support for equity theory in two significant ways. First, they indicate that the predictions of equity theory about the effects of payment on productivity which previously had been primarily tested in overpayment situations also appear to hold in an underpayment situation. Second, the results demonstrate, as predicted by equity theory, that in addition to using changes in productivity and work quality to reduce dissonance, *S*s also use attitudinal and perceptual mechanisms to arrive at an equitable input-outcome balance.

The personality measures and the measure of the importance of the money earned to the *S* were included in order to understand what importance these factors might have in determining the degree to which *S*s used the different dissonance-reduction mechanisms. Perhaps the most significant result here is the finding that the importance of the money to the *S* is significantly associated with *S*'s trying to increase his productivity in order to reduce dissonance. Not surprisingly, *S*s who need money more are more likely to increase their productivity. This finding is quite predictable from Vroom's (1964) theory of motivation and from other path-goal theories, but it is a good example of the kind of prediction, about which dissonance-reduction mechanism will be chosen by an individual, that equity theory frequently fails to make.

The data from the personality test provide some clues as to the types of individuals who are likely to raise their productivity in order to be able to reduce their dissonance. Interestingly, those *S*s who are low on the measures of poise, ascendancy, and self-assurance seem to be characterized by high productivity. Weick (1965) has suggested that one effect of low reward is that it raises a person's concern about his worthiness, competence, and adequacy. Clearly this would be expected to happen much more easily in *S*s who have low self-assurance already. For these *S*s it may be that raising productivity presents the only concrete way of reducing these feelings of guilt and self-doubt, and showing the *E* and themselves that they really are worthwhile. Task enhancement and input reduction might not be effective for this personality type. To the extent that a large number of *S*s in the underpaid group were motivated by a self-confidence threat rather than by simple dissonance, this may offer an alternative interpretation to the impact of underpayment on productivity. However, it does not appear to explain the attitude differences as well as does the equity-theory explanation.

The maturity and responsibility scales tend to correlate significantly with work quality. Apparently for people with a high sense of responsibility and maturity, sacrificing quality for quantity was an untenable solution to the dissonance situation. It might even be argued that for these *S*s the dissonance was not great because it failed to threaten their perceptions of self-worth. Independent of this, however, it probably is generally true that *S*s with a high sense of responsibility will be less likely to choose a dissonance-reduction means that

forces them to reduce quality than will *Ss* who are low on responsibility.

It is surprising that the task-enhancement attitudes and the job-input attitudes failed to present any consistent relationship to the personality factors considered. Still, the results of the personality test do argue that personality factors are important in determining which mechanism will be chosen and that they must be considered in any comprehensive theory of dissonance reduction. In summary, it appears that equity theory did a good job of predicting the inequity-reduction mechanisms that would be used by the underpaid *Ss*. However, when it came to understanding the role of personality and the importance of outcomes, it appears that further elaboration is needed if the theory is to handle individual differences in reactions to an inequity situation.

REFERENCES

ADAMS, J. S. Toward an understanding of inequity. *Journal of Abnormal and Social Psychology*, 1963, 67, 422–436. (a)

ADAMS, J. S. Wage inequities, productivity and work quality. *Industrial Relations*, 1963, 3, 9–16. (b)

ADAMS, J. S. Injustice in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 2. New York: Academic Press, 1965. Pp. 267–299.

ADAMS, J. S., & JACOBSEN, P. R. Effects of wage inequities on work quality. *Journal of Abnormal and Social Psychology*, 1964, 69, 19–25.

ADAMS, J. S., & ROSENBAUM, W. E. The relationship of worker productivity to cognitive dissonance about wage inequities. *Journal of Applied Psychology*, 1962, 46, 161–164.

ANDREWS, I. R. Job performance as a function of wage inequity. Unpublished doctoral dissertation, University of California, 1965.

ARROWOOD, A. J. Some effects on productivity of justified and unjustified levels of reward under public and private conditions. Unpublished doctoral dissertation, University of Minnesota, 1961.

BLAU, P. *Exchange and power in social life*. New York: Wiley, 1964.

FESTINGER, L. *A theory of cognitive dissonance*. Evanston, Ill.: Row, Peterson, 1957.

GOUGH, H. G. *California Personality Inventory*. Palo Alto: Consulting Psychologists Press, 1957.

HOMANS, G. C. *Social behavior: Its elementary forms*. New York: Harcourt, Brace & World, 1961.

VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.

WEICK, K. E. Reduction of cognitive dissonance through task enhancement and effort expenditure. *Journal of Abnormal and Social Psychology*, 1964, 68, 533–539.

WEICK, K. E. The concept of equity in the perception of pay. Paper read at a meeting of the Midwestern Psychological Association, Chicago, April 1965.

(Received October 7, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 411–416

EFFECTS OF MONETARY REWARDS AND PUNISHMENTS ON VIGILANCE PERFORMANCE[1]

RUSSELL L. SMITH,[2] LUIGI F. LUCACCINI, AND MURRAY H. EPSTEIN

University of California, Los Angeles

In a complex visual vigilance task lasting 1 hr., Ss in 5 experimental groups were rewarded for correct detections and punished either for missed targets or for false alarms. Ss in a control group performed the task without possibility of reward or punishment. 3 levels of monetary incentive were used as rewards. The major results of the study indicated that (a) some combinations of reward and punishment facilitated detection performance while others did not, and (b) Ss punished for missed targets performed better than Ss punished for false alarms. Implications for vigilance research and theory are discussed.

The results of most vigilance research can be separated into two broad categories, depending on the type of display used. With simple displays, such as the Mackworth Clock Test (Mackworth, 1950), the percentage of correct detections declines rapidly from an initially high level to a relatively low and thereafter constant level. This decline is the well-known vigilance decrement. With more complex displays, such as multiple-dial monitoring (Jerison & Wallis, 1957), performance tends to be low and constant throughout the task, and decrements are seldom found.

Although much of vigilance research has been concerned with discovering what physical-task parameters may be responsible for the low level of performance found with both kinds of displays, some investigators feel that motivation is the key to understanding vigilance behavior. Mackworth (1961), for example, has pointed out that the performance of "dedicated" *Ss* far outstrips that of the typical *S*. Moreover, Elliott (1960) has argued that the vigilance decrement does not even occur in military monitoring tasks, and Kibler (1965) has doubted the relevance of laboratory vigilance results in general. Unless one is willing to accept the questionable assumption that laboratory *Ss* are as concerned in vigilance tasks as, for example, radar operators, the proposition must be accepted that "typical" vigilance results may be, in large part, artifactual. Rejection of that assumption logically suggests a more rigorous evaluation of the effects of motivation. Existing theories of vigilance, save one (Smith, 1966), have not treated the variable of motivation to any significant degree. In view of some rather dramatic results obtained by early investigations (Fraser, 1953; Mackworth, 1950) which employed certain modes of motivation, the failure to give sufficient consideration to that variable is somewhat surprising.

The present study attempts to investigate the effects of motivation on vigilance performance by the use of monetary incentives. Three prior studies (Bergum & Lehr, 1964; Pollack & Knaff, 1958; Sipowicz, Ware, & Baker, 1962) have employed monetary incentives with different results. Pollack and Knaff (1958) reported a slight (and apparently nonsignificant) improvement in detection performance when *Ss* were offered "an extra hour of pay" for "either 100% detection in seven of eight 10-minute periods or the greatest improvement in scores over the previous tests." Since *Ss* were already guaranteed payment for 50 hr. of service and since the reward was to some extent competitive, it is questionable whether the offer was an effective inducement.

Sipowicz et al. (1962) found monetary incentives to be of considerable benefit, while Bergum and Lehr (1964) did not. In the latter study the performance of rewarded *Ss*

[1] The authors wish to express their appreciation to the Department of Psychology at the University of California, Los Angeles for providing funds for this study, and to John Lyman for providing experimental facilities.

[2] Also at Dunlap and Associates, Inc., Santa Monica.

was superior to that of unrewarded Controls only for the first of three consecutive 20-min. periods. Both studies employed similar displays, small amounts of money, and similar *S* populations (army recruits). Moreover, *S*s in both studies were punished for missed targets by a reduction in the total earned for correct detections, although more severely so in the former study.

The current study attempts (*a*) to extend the findings of Sipowicz et al. (1962) to complex displays, and (*b*) to evaluate the effects of different levels of reward and modes of punishment on vigilance performance. Five combinations of reward and punishment were used. In one punishment condition, *S*s were penalized for missed targets, while in the other, for false alarms. Hypotheses derived from the theoretical model of Smith (1966) were: (1) detection performance of rewarded *S*s would be facilitated over that of Controls in proportion to the amount of reward offered; (2) at each level of reward, detection performance would be better for *S*s punished for missed targets; and (3) at each level of reward, false-alarm rates would be higher for *S*s punished for missed targets.

Method

Subjects. The *S*s were 24 male and 24 female undergraduates from introductory psychology courses at the University of California at Los Angeles. Each *S* received credit toward a course requirement for participating in a single test session of 90 min. duration. Assignment of *S*s to conditions within the experiment was made randomly.

Apparatus. The visual display employed was identical to that used in another study (Smith, Lucaccini, Groth, & Lyman, 1966) in which it is described in greater detail. The display was designed to simulate a complex, constantly changing, visual search and discrimination problem. The object of the task was to observe a small window (2 in. high × 0.5 in. wide) and report the appearance of a target by pressing a response button. The display window was located at eye level in the center of a U-shaped isolation booth. A constant-speed motor pulled a white paper tape from left to right behind the window. Three types of solid geometric objects appeared on the tape: targets, false targets, and a randomappearing visual "noise" background. Targets and false targets were small black squares that differed in the location of a small extension. Objects in the visual noise background were also small black squares that differed from targets and false targets in the number of sides containing extensions. Sixty minutes were required to present the entire tape, which moved at a speed of 0.5 in./sec.

For purposes of target presentation the tape was divided into six consecutive 10-min. display units. Within each unit, four targets occurred at randomly selected temporal intervals. Two targets each appeared alone against the noise background; the other two appeared with a false target in the opposite half of the display. Two more false targets appeared at random temporal intervals within the display unit. Thus, over the entire experiment 24 targets and 24 false targets appeared. Twelve of each appeared alone against the noise background and 12 appeared as target-false-target pairs.

Experimental conditions. The experiment was basically a 6 × 2 × 2 × 3 repeated-measures design with six incentive conditions, two sexes, two target types, and three time intervals. Independent groups of 8 and 24 *S*s served in each incentive and sex condition, respectively. The remaining variables were repeated measures taken on each *S*.

Incentives. Eight *S*s served in each incentive condition.

Control (Condition C): *S*s were told to observe the display continually and report targets as soon as they were detected.

Low Incentive, Errors Punished (Condition LE): *S*s were instructed as in Condition C. In addition, they were told that each correct detection would be rewarded with a payment of 10¢ at the end of the experiment. Every erroneous report (false alarm) would be punished by a deduction of 10¢ from the amount earned. No punishment was given for missed targets.

Low Incentive, Misses Punished (Condition LM): *S*s received similar instructions to Condition LE except that 10¢ was deducted for missed targets while false alarms were not penalized.

Medium Incentive, Errors Punished (Condition ME): Similar to Condition LE except that rewards and penalties consisted of 20¢.

Medium Incentive, Misses Punished (Condition MM): Similar to Condition LM except that rewards and penalties consisted of 20¢.

High Incentive, Errors Punished (Condition HE): *S*s were instructed as in Condition C. In addition, they were informed that they would be given "a certain fixed amount of money" for each correct detection upon completion of the experiment. False alarms would be penalized by subtracting one-half of the reward for a correct detection from the total amount earned. Missed targets would not be penalized. It was possible to earn "up to $10 for perfect performance and $4 or $5 for average performance." (Condition HE was included to induce a very high level of motivation. When it became apparent early in the study that Condition HE *S*s indicated doubt regarding the genuineness of the reward, a balancing condition, HM, was dropped for reasons of economy.)

Sex. Each incentive group was composed of four males and four females to assess the effects of sex

of *S* on performance. This variable was included to provide another test of an earlier finding that there were no differences in performance between males and females (Smith et al., 1966).

Target type. This variable reflected the attempt to vary the level of task difficulty by presenting targets alone or with false targets against the noise background. Targets presented alone have been shown to be detected more easily (Smith et al., 1966).

Time interval. This variable corresponded to the division of the task into three successive 20-min. periods in order to assess sequential effects.

Procedures. The *Ss* served individually in the task. General instructions were read to each *S* at the beginning of the experiment explaining the nature of the task and the response procedures. A 2-min. practice session followed in which knowledge of results was given verbally by *E* after each target or false target appeared. Further instructions were then given to *Ss* in the various incentive conditions, questions about procedure were answered, and the task began. No interruptions were permitted during the hour. Communication with *E*, who remained behind the display, was prohibited, and *S* was asked to remove his watch. At the end of the hour, payment of the sum earned was given to *Ss* of incentive groups. A questionnaire was then administered on which *Ss* estimated the duration of the task in minutes, reported the amount of money they had expected to earn, and rated the following items: (*a*) interest in the detection task, on a 5-point scale ranging from "very interesting" to "very boring," and (*b*) belief at start of the experiment in the genuineness of the incentive offer, on a 5-point scale ranging from "definitely thought I would be paid" to "definitely thought I wouldn't be paid."

Performance measures. Two separate performance measures were recorded for each *S*. These were the number of correct target detections and the number of false alarms (incorrect reports that a target had appeared). These two measures were derived from recordings of target presentations and *Ss'* responses (button presses) made with a four-channel Gerbrands event recorder.

RESULTS

The number of correct detections was summed for each *S* and tested for significant differences between conditions by analysis of variance. The results showed that two main effects reached significance: Incentive Condition—F $(5/36) = 2.67$, $p < .05$—and Target Type—F $(1/36) = 9.09$, $p < .01$. The only first-order interaction to reach significance was Target Type × Time Period—F $(2/72) = 9.93$, $p < .001$. One higher-order interaction was significant, Incentive Condition × Sex × Time Period × Target Type—F $(10/72) = 2.22$, $p < .05$.

TABLE 1

MEAN PERCENTAGE DETECTIONS BY TARGET TYPE AND TIME PERIOD

Target presentation mode	Successive 20-min. intervals		
	1	2	3
Alone	68.2	60.5	64.0
With false target	45.8	56.2	65.5

Figure 1 presents mean detection performance by successive 20-min. periods under each of the six incentive conditions. A curve representing performance in a previous study (Smith et al., 1966) is included in the figure for comparison and will be discussed below. A Newman-Keuls test (Winer, 1962) of the overall means for the six incentive conditions revealed that detection performance was significantly higher ($p < .01$) in Conditions LM and MM than in the other four conditions, with Condition MM superior to Condition LM ($p < .01$). Differences between the remaining four conditions did not reach significance.

Table 1 presents mean detection performance as a function of target type and time period. A Newman-Keuls test of the means for this interaction revealed that, although overall performance was better for targets presented alone, the difference in performance between the two modes of target presentation was significant ($p < .01$) only for the initial 20-min. period of the task. Performance did not differ significantly among the three intervals for targets presented alone nor between the last two intervals for targets presented with a false target.

Analysis of variance failed to reveal significant differences between incentive conditions in the number of false alarms committed.

Analysis of responses to the postexperimental questionnaire revealed significant differences for the question regarding *S*'s belief in the genuineness of the incentive offer ($\chi^2 = 13.33$, $df = 1$, $p < .001$). Six of the eight *Ss* in Condition HE stated that when they started the task they did not expect to receive the promised payment upon completing the task. In contrast, 28 of the 32 *Ss* in Con-

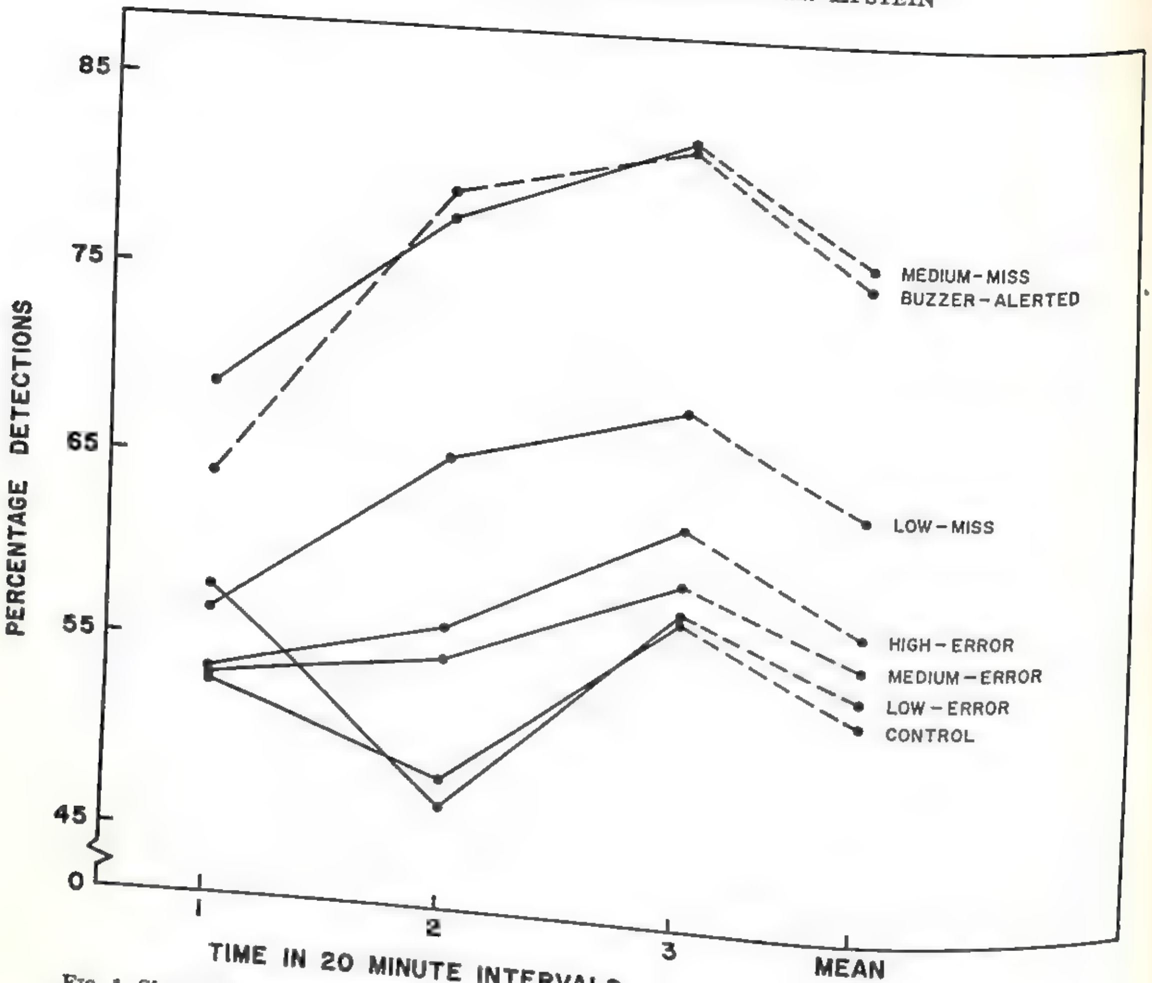


Fig. 1. Signal detection performance under various incentive conditions as a function of time interval. The completely dashed curve (buzzer-alerted) represents data taken from Smith et al. (1966).

ditions LE, LM, ME, and MM felt that they would "probably" or "definitely" be paid. The *Ss* in Condition C were not asked this question.

Chi-square analysis of responses to the interest question revealed that the five rewarded groups (LE, LM, ME, MM, and HE) all rated the task as significantly less boring than the control group ($\chi^2 = 5.76$, $df = 1$, $p < .02$). The *Ss* in Condition C rated the task as boring, while all other groups rated it at about the midpoint ("so-so") of the interest scale.

Analysis of variance of the mean earnings estimates for the five rewarded incentive groups revealed significant differences between groups—F $(4/35) = 3.30$, $p < .05$. A Newman-Keuls test showed that the mean estimate for Condition HE ($3.94) was significantly ($p < .05$) higher than for the other four groups. Differences between the other four groups did not reach significance. Mean estimates for Conditions LE, LM, ME, and MM were $1.28, $1.43, $1.26, and $2.06, respectively.

There were no significant differences between the six incentive conditions in estimates of task duration. All groups underestimated the duration of the vigilance task ($\bar{X} = 44.9$ min.).

Discussion

The results of this study provide support for the hypothesis that when motivation is

increased by monetary incentives, performance on a complex vigilance task can be significantly enhanced. Although only two of the five rewarded incentive groups (Conditions LM and MM) performed significantly better than Controls, there is evidence that the motivational technique employed was not effective in the other three rewarded groups. In the case of Condition HE, it is clear from answers to the postexperimental questionnaire that, although earnings estimates were highest for this group, *S*s in this group were nearly unanimous in their belief that the money earned during the task would not be received upon completion. In Conditions LE and ME, although *S*s apparently accepted the incentive offer as genuine, the method of punishment (for false alarms) may have seemed so severe as to reduce the chance of earning a large amount of money. Answers to the postexperimental questionnaire tend to support this interpretation; *S*s in these two groups reported lower earning estimates than *S*s in Conditions LM and MM.

The significance of the performance enhancement found in Condition MM on this complex vigilance task can be appreciated by comparison with the performance of an alerted group from a prior study that employed similar *S*s and the same vigilance task (Smith et al., 1966). In that study, uncertainty of signal presentation was eliminated for the alerted *S*s by sounding a buzzer 1 sec. before the signal appeared. Figure 1 shows that the performances of the two groups were nearly identical.

The hypothesis that, at each level of reward, *S*s punished only for missed targets would detect more targets than those punished only for false alarms was clearly supported. In fact, the effects of method of punishment were so strong as to overshadow the effects of level of incentive. For example, *S*s paid 10¢ per detection and punished for missed targets (Condition LM) outperformed *S*s paid 20¢ per detection but punished for false alarms (Condition ME).

The failure to find differences in detection performance between male and female *S*s is consistent with the results of an earlier study (Smith et al., 1966) employing the same vigilance display. This finding is in contrast to that of Whittenburg, Ross, and Andrews (1956) who found the performance of females to be superior to that of males on the second half of a 2-hr. task.

The significant improvement in detection performance that occurred for the more difficult target type (i.e., those presented with a false target) is in agreement with the results of the earlier study (Smith et al., 1966) using these two target types. In both studies the most likely explanation is that some learning occurred during the initial part of the task.

Task-interest ratings indicated that the vigilance task may have been less boring for *S*s in the five rewarded groups than for Controls. Nevertheless, all groups indicated some degree of boredom. This finding suggests that the essentially monotonous nature of the vigilance task was not altered by incentives even though performance was facilitated.

Finally, what explanation can be offered for the fact that rewards were beneficial in the present study (Conditions LM and MM) and in that of Sipowicz et al. (1962), but of only temporary benefit for Bergum and Lehr (1964)? In all three investigations rewards were earned in proportion to the number of targets detected, earnings were decreased for missed targets, and the maximum possible earnings were about the same. In the Sipowicz et al. (1962) study *S*s were informed of the maximum possible reward ($3.00), but in the current study, with the exception of Condition HE, and in that of Bergum and Lehr (1964) *S*s were informed only of the reward per signal (10 or 20¢) and of the cost of errors. In the latter two studies, *S*s had to have some idea of the signal rate in order to estimate whether or not the task was "worthwhile" in terms of probable earnings. This knowledge could be gained only after serving in the study for some period of time. By the end of the initial 20-min. period of vigilance, *S*s in Bergum and Lehr's (1964) study may have realized that the signal rate was low (12/hr) and decided the potential earnings did not warrant continued "extra effort." This would explain why the performance of their rewarded group subsequently dropped to the level of the Controls. In the present

study, with a signal rate of 24/hr, *S*s in Conditions LM and MM apparently decided that "extra effort" would be adequately compensated.

The results of the present study are generally consistent with the predictions generated from Smith's (1966) motivational theory of vigilance and add to a growing body of evidence which suggests that the low level of vigilance performance typically found in the laboratory may be an artifact. The current study and that of Sipowicz et al. (1962) suggest that vigilance performance may be significantly enhanced when *S*s are reasonably rewarded for increased vigilance. Punishment (Pollack & Knaff, 1958), coercion (Bergum & Lehr, 1963; Fraser, 1953), and knowledge of results—true, false, or irrelevant (Baker, 1961; Hardesty, Trumbo, & Bevan, 1963; Loeb & Schmidt, 1960; Mackworth, 1950; McCormack, 1959; Sipowicz et al., 1962; Weidenfeller, Baker, & Ware, 1962)—have also been shown to enhance vigilance performance when properly applied. None of those studies involved a change in the physical parameters of the vigilance task itself. Taken together they indicate that a truly comprehensive theoretical account of vigilance behavior must encompass motivational variables.

REFERENCES

BAKER, C. H. Maintaining the level of vigilance by means of knowledge of results about a secondary vigilance task. *Ergonomics*, 1961, **4**, 311–316.

BERGUM, B. O., & LEHR, D. J. Effects of authoritarianism on vigilance performance. *Journal of Applied Psychology*, 1963, **47**, 75–77.

BERGUM, B. O., & LEHR, D. J. Monetary incentives and vigilance. *Journal of Experimental Psychology*, 1964, **67**, 197–198.

ELLIOTT, E. Perception and alertness. *Ergonomics*, 1960, **3**, 357–364.

FRASER, D. C. The relation of an environmental variable to performance in a prolonged visual task. *Quarterly Journal of Experimental Psychology*, 1953, **5**, 31–34.

HARDESTY, D., TRUMBO, D., & BEVAN, W. Influence of knowledge of results on performance in a monitoring task. *Perceptual and Motor Skills*, 1963, **16**, 629–634.

JERISON, H. J., & WALLIS, R. A. Experiments on vigilance: One-clock and three-clock monitoring. *United States Air Force Wright Air Development Center Technical Report*, 1957, No. 57-206.

KIBLER, A. W. The relevance of vigilance research to aerospace monitoring tasks. *Human Factors*, 1965, **7**, 93–99.

LOEB, M., & SCHMIDT, E. A. Influence of time on task and false information on efficiency of responding to pure tones. *United States Army Medical Research Laboratory Report*, 1960, No. 426.

MACKWORTH, N. H. Researches on the measurement of human performance. *Medical Research Council Special Report Series*, 1950, No. 268.

MACKWORTH, N. H. Human vigilance. *Electro-Technology*, May 1961, 121–123.

MCCORMACK, P. D. Performance in a vigilance task with and without knowledge of results. *Canadian Journal of Psychology*, 1959, **13**, 68–71.

POLLACK, I., & KNAFF, P. R. Maintenance of alertness by a loud auditory signal. *Journal of the Acoustical Society of America*, 1958, **30**, 1013–1016.

SIPOWICZ, R. R., WARE, J. R., & BAKER, R. A. The effects of reward and knowledge of results on the performance of a simple vigilance task. *Journal of Experimental Psychology*, 1962, **64**, 58–61.

SMITH, R. L. *Monotony and motivation: A theory of vigilance*. Santa Monica: Dunlap and Associates, Inc., 1966.

SMITH, R. L., LUCACCINI, L. F., GROTH, H., & LYMAN, J. The effects of anticipatory alerting signals and a compatible secondary task on vigilance performance. *Journal of Applied Psychology*, 1966, **50**, 240–246.

WEIDENFELLER, E. W., BAKER, R. A., & WARE, J. R. The effects of knowledge of results (true and false) on vigilance performance. *Perceptual and Motor Skills*, 1962, **14**, 211–215.

WHITTENBURG, J. A., ROSS, S., & ANDREWS, T. G. Sustained perceptual efficiency as measured by the Mackworth "clock" test. *Perceptual and Motor Skills*, 1956, **6**, 109–116.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received October 7, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 417–425

SEARCHING FOR NEWSPAPER HEADLINES PRINTED IN CAPITALS OR LOWER-CASE LETTERS [1]

E. C. POULTON

Applied Psychology Research Unit, Cambridge, England

Altogether 264 adults searched for headlines which were printed 3 times in various members of the Times group of typefaces. There were 2 newssheets, and a 2 × 2 factorial design was used to compare pairs of alternative printings. Headlines printed in Titling or Extended Titling capitals with heights of about 4.5, 4.0, 3.0, and 2.5 mm. according to their importance, took reliably ($p < .01$) longer to locate than when printed in bold lower-case letters whose x-heights approximately matched the heights of the capital letters. Subsidiary headlines printed in smaller letters above or below the main headlines distracted attention from the main headlines ($p < .05$). When most of the main headlines were printed in lower-case letters, the subsidiary headlines were more likely to attract attention when they were printed in capitals than when printed in lower-case letters of the same point size ($p < .05$).

Most people who take a daily newspaper do not read it systematically like a fiction paperback, starting at the top of the first page and ending at the bottom of the last page. Instead they glance at the headlines, and read only the paragraphs of news below the headlines which look interesting. Even the paragraphs of news are not always read systematically when the reader is in a hurry. He may simply run his eyes down a paragraph looking for news of a particular sort. Common observations of this kind on readers of newspapers suggest that the criteria which have been used in the past to evaluate experimentally the printing and layout of newspapers may not have been the most appropriate.

When investigating the printing of newspaper headlines, Paterson and Tinker's (1946) research students presented one headline at a time for a limited period. Their measure was the number of words read in the time available. At a normal reading distance, reliably more words were read on the average when using two-line single-column headlines printed in 24-point Cheltenham Extra Condensed lower-case letters than when using headlines printed all in capitals (upper-case letters) of the same type face and point size. Sixty-point Memphis Bold capitals were at a reliable advantage over the corresponding lower-case letters in multicolumn headlines only at a reading distance of 17 ft., where the greater average height of the capital letters made them more visible.

Unfortunately this rate-of-reading method measures only the readability of individual headlines. It is not a measure of the attention-getting value of a headline located on a sheet of newspaper, which is what newspaper editors are (or should be) interested in. Attention getting is related to the discriminability of the headline from the other headlines on the page, from the advertisements, the pictures, and the columns of news. It is measured more validly by presenting the headline in its setting in a newssheet, and by determining how long the average reader takes to find it. This is the basis of the method used in the series of experiments described here.

There is at present no acceptable way of equating upper-case and lower-case letters. Point size cannot be used, because it specifies only the height of the block upon which the letter is cast for printing. It does not specify the width of the block, which determines the number of letters printed in a line of fixed

[1] This research was carried out at the request of Sir William Haley, then editor of The Times. The author is grateful to the Editor for providing the pages of newspaper, to Walter James, editor of The Times Educational Supplement, for his help and encouragement, and to Walter Tracy, typographical adviser to The Times Publishing Company, for specifying the typographical arrangements used. P. M. E. Altham kindly advised on the statistics. Financial support from the British Medical Research Council is also gratefully acknowledged.

length. Upper-case letters are normally wider than lower-case letters of the same point size, and require a line about 35% longer (Tinker, 1963, p. 60). Perhaps less well known to applied psychologists, point size does not specify exactly the height of the face; the relationship may vary with the typeface. For example, the capitals of the top headlines of Figure 1a and b have practically identical heights, yet the point sizes are 14 and 18, respectively. Similarly lower-case letters look large for their point size if the x-height (the height of the rounded parts of the letters) is large in proportion to the heights of the ascenders and descenders (see Poulton, 1965).

If lines of upper-case letters are printed too close together, the shapes of the tops and bottoms of the letters are masked by the adjacent letters above and below. The same can happen to lines of lower-case letters if they are designed with the x-height taking up a disproportionate amount of the point size. The masking between lines of capitals can be prevented either by holding the height of the face constant while increasing the point size, or by leading between the lines. Similarly the masking between lines of lower-case letters can be prevented either by holding the x-height constant while increasing the heights of the ascenders and descenders, or again by leading. When the x-height of lower-case letters is made to match the overall height of capitals, names printed in lower-case letters with capitals only for first letters are visible at rather greater distances than names printed all in capitals (Forbes, Moscowitz, & Morgan, 1950). When names arranged one below the other are optimally spaced and equated for size in terms of area covered, the x-height of the lower-case names is rather smaller than the overall height of the names in capitals, and they cannot be read from quite so far (Moore & Christie, 1963, p. 116).

To the author's knowledge no tests have been carried out to see if printed upper- and lower-case words are equally readable or discriminable at normal reading distances when equated in this way. However this was not the aim of the present experiments. The aim here was to compare an existing style of upper-case headlines which had hardly been changed since it was first introduced in 1932, with new designs of upper- and lower-case headlines. Usually the overall heights of the large- and medium-sized upper-case letters matched the x-heights of the corresponding lower-case letters, as illustrated in Figure 1. The vertical spacing between the lines of upper-case letters was only about half the vertical spacing between the rounded parts of the lower-case letters. The words set in large and medium upper-case letters were usually about 5% longer than the corresponding lower-case words.

Method

Materials

The newspaper was laid out in seven columns containing lines of print 2.2 in. long. The printed pages were 16.3 in. wide and 22.0 in. tall. One of the pages carried an advertisement which occupied the full length of the three left-hand columns. The layout of the remaining four columns is illustrated in Figure 2. The complete bottom half of the other page was taken up by an advertisement which extended across all seven columns. The layout of the top half of the page is illustrated in Figure 3.

All the headlines in the two original pages had been printed in capitals, as in Figures 2a and 3a. The two pages were reprinted twice, once with all the headlines set in lower-case lettering, capitals being used only for the first letters of headlines and for proper names, as in Figures 2c and 3c. On the second reprinting half the headlines were set in capitals and half in lower case, as in Figures 2b and 3b. All the headlines were printed in various members of the Times group of typefaces.

In the original pages, six of the eight large *upper-case* headlines had letters about 4.0 mm. tall. The double-column headline of Figure 2a and the double-column headlines in the middle and on the right of Figure 3a were set in 14-point Extended Titling Monotype Series Number 339 with 6-point leading, which is illustrated actual size by the top headline of the pair in Figure 1a. The two single-column headlines at the top of Figure 2a and the one to the left of the weather map in Figure 3a were set in 18-point Titling Monotype Series Number 329 with 3-point leading, which is illustrated actual size by the top of the pair in Figure 1b. The treble-column headline toward the bottom of Figure 2a and the double-column headline on the left of Figure 3a were set in 16-point 339 which has letters about 4.5 mm. tall. The double-column headline had 9-point leading. Six of the large headlines were paired with subsidiary headlines. Four "second-deck" subsidiaries with letters about 3.0 mm. tall were located immediately below the main headlines. The double-column subsidiary of Figure 2a and the right-hand double-column subsidiary of Figure 3a were printed in 12-point 339 with 3-point leading, which is illustrated actual size in

a

ZAMBIA ASKS AFRICAN STATES FOR ECONOMIC HELP

O.A.U. CONFERENCE EXPECTED TO SET UP SPECIAL FUND

ADDIS ABABA, MARCH 2 | the British members of the board of trustees

b

AUSTRIAN PRESS FREEDOM CLASH

COURT ORDERS TO EDITOR

FROM OUR OWN CORRESPONDENT

VIENNA, MARCH 2

With only two days to go to the Aus-

c

NEW TANK FOR EAST GERMAN ARMY

FROM OUR CORRESPONDENT

BERLIN, MARCH 2

The east German Army has been receiving a new tank, the T55, since the beginning of the year to replace the T54. General Hoffmann, the Defence Minister,

d

FIRST TRIDENT FOR PAKISTAN

Group Capt. John Cunningham, chief test pilot of the British Aircraft Corporation, piloted the first three-jet Trident airliner of Pakistan International Airways

e

ZAMBIA SEEKS ECONOMIC HELP

African conference expected to set up a fund

ADDIS ABABA, MARCH 2 | the British members of the board of trustees

f

Zambia seeks economic help

African conference expected to set up a fund

ADDIS ABABA, MARCH 2 | the British members of the board of trustees

g

New tank for east German army

FROM OUR CORRESPONDENT

BERLIN, MARCH 2

h

First Trident for Pakistan

Group Capt. John Cunningham, chief test pilot of the British Aircraft Corporation, piloted the first three-jet Trident airliner of Pakistan International Airways

FIG. 1. Full-size examples of the headlines used. (a, b, c, and d are illustrations from the newssheets with upper-case headlines. e illustrates the larger paired headlines from the newssheets with mixed upper- and lower-case headlines. Two-thirds of the smaller headlines were set in lower case as in g, the remaining third in upper case as in c. All the small headlines were set in upper case as in d. f, g, and h are illustrations from the newssheets with lower-case headlines.)

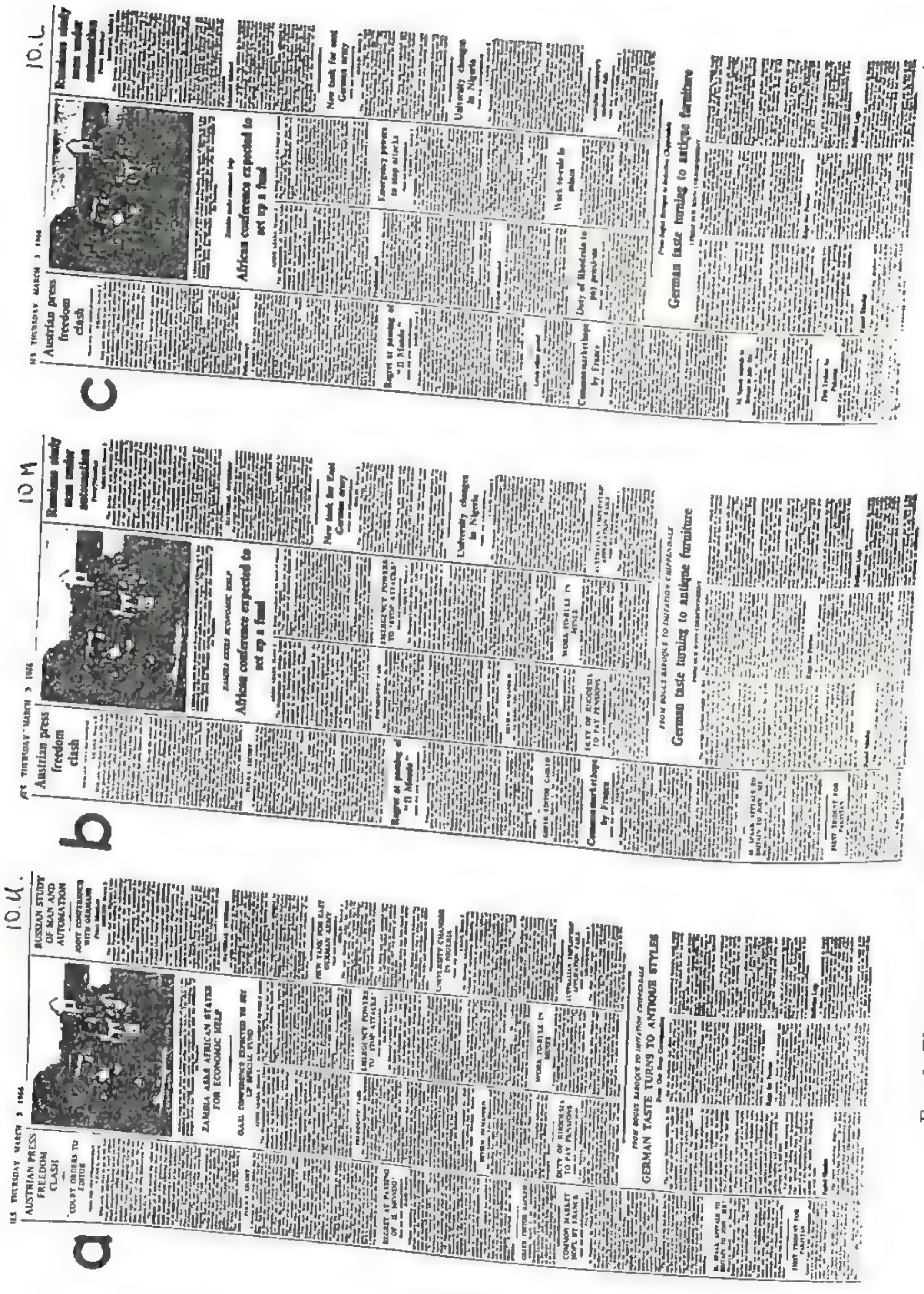


FIG. 2. The news columns of one of the pages. (A pictorial advertisement which occupied the full length of the three left-hand columns is not shown. a. All headlines in upper case. b. Mixed upper- and lower-case headlines. c. All headlines in lower case.)

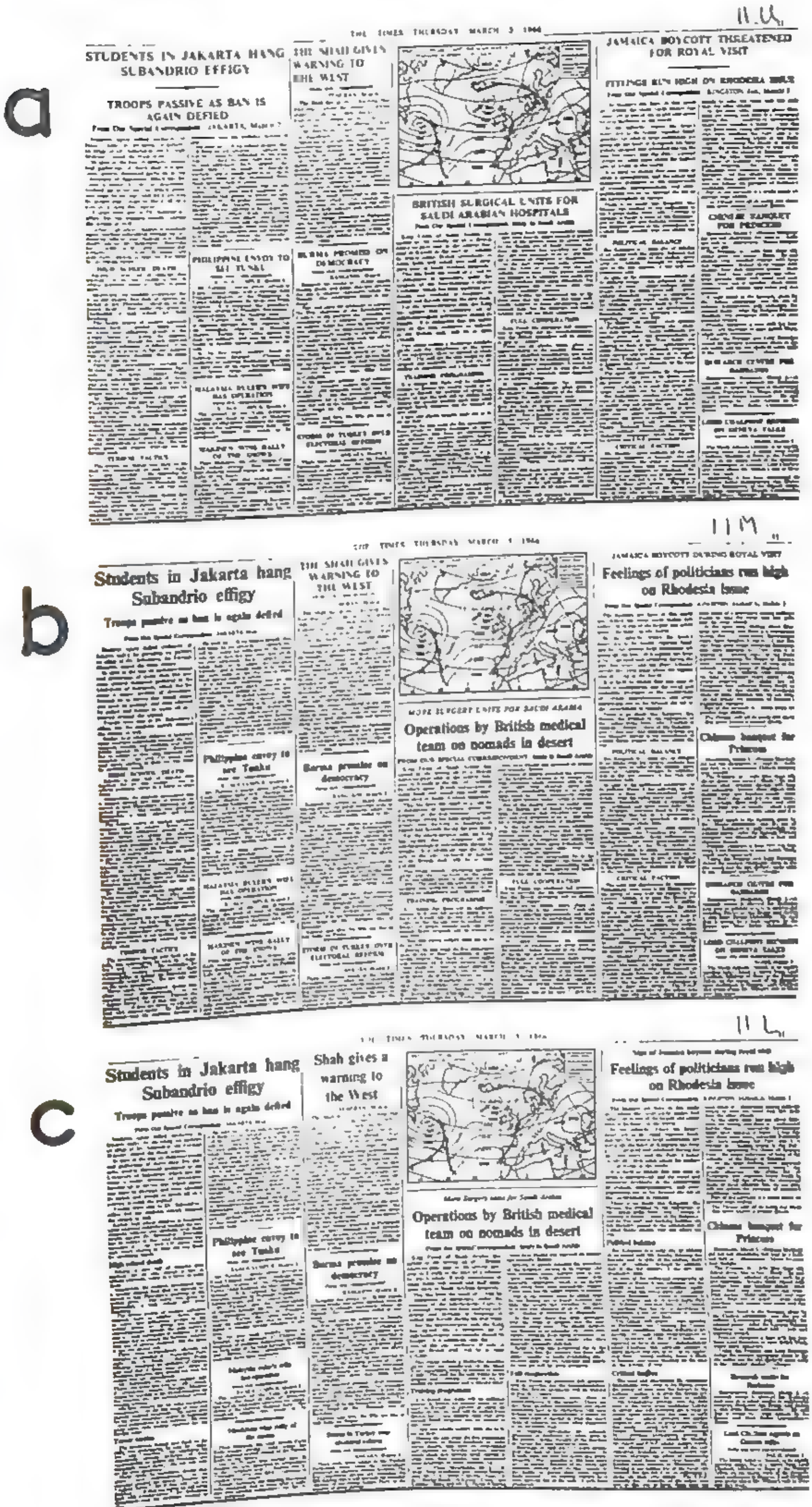


FIG. 3. The news columns of the other page. (A pictorial advertisement which occupied the complete bottom half of the page is not shown. a. All headlines in upper case. b. Mixed upper- and lower-case headlines. c. All headlines in lower case.)

Figure 1a. The two single-column subsidiaries of Figure 2a were printed in 14-point 329 with 1.5-point leading, as in Figure 1b. The large double-column headline on the left of Figure 3a was paired with a second-deck subsidiary beneath it in 14-point 339 with 6-point leading. The large treble-column headline toward the bottom of Figure 2a was paired with a "strap-line" subsidiary printed above it in 12-point italic capitals about 2.5 mm. tall of Times Roman Monotype Series No. 327. The medium-sized unpaired headlines restricted to a single column had letters about 3.0 mm. tall. They were set in 12-point 339 with 3-point leading as illustrated actual size in Figure 1c. The small unpaired headlines restricted to a single column had letters about 2.5 mm. tall. They were set in 10-point 339 with 3-point leading, as in Figure 1d.

In the reprinted pages, seven of the eight large *lower-case* headlines were set in letters with an x-height of about 4.0 mm., using 24-point Times Bold Monotype Series No. 334. The three large headlines at or toward the top of Figure 2c and the double-column headlines in the center and on the right of Figure 3c had 6-point leading, as illustrated actual size by the bottom headline of Figure 1f. The single-column headline to the left of the weather map in Figure 3c had 9-point leading. The eighth large headline on the left of Figure 3c had an x-height of about 5.0 mm. It was printed in 30-point 334 with 6-point leading. Five of the large headlines were paired with subsidiary headlines. The double-column and treble-column headlines of Figure 2c and the double-column headlines in the center and on the right of Figure 3c had strap lines printed above them in letters with an x-height of about 1.5 mm., using 12-point Times Semi-bold Monotype Series No. 421. Three of the strap lines were in italics, as in Figure 1f. The fourth on the right of Figure 3c was printed in roman letters. The large double-column headline on the left of Figure 3c had a second-deck subsidiary beneath it with an x-height of about 3.0 mm., set in 18-point 334. The medium-sized unpaired headlines restricted to a single column were also set in 18-point 334, half with 4.5-point leading, half with 3-point leading as in Figure 1g. The small unpaired headlines restricted to a single column had an x-height of about 2.0 mm. They were set in 12-point 334 with 3-point leading, as in Figure 1h.

Seven out of the eight large headlines from the two newssheets with *mixed upper- and lower-case* headlines were identical with the corresponding headlines in the lower-case headline newssheets. The eighth, to the left of the weather map in Figure 3b, was set in 14-point 339 capitals with 6-point leading like the top headline in Figure 1a. The five subsidiary headlines corresponded to the five in the newssheets with lower-case headlines. The four strap lines were printed in capital letters about 2.5 mm. high, using 12-point 421. As before, three of the strap lines were in italics, as in Figure 1e. The fourth, on the right of Figure 3b, was printed in roman letters. The fifth headline pair, on the left of Figure 3b, was identical with the pair in the lower-case-headline newssheet. Of the medium-sized unpaired headlines restricted to a single column, two-thirds were identical with headlines in the lower-case-headline newssheets (Figure 1g). The remaining third were identical with headlines in the upper-case-headline newssheets (Figure 1c). The small unpaired headlines restricted to a single column were all identical with the upper-case headlines (Figure 1d).

No attempt was made in the reprinted pages to standardize style of type or point size at the cost of what the typographic designer considered to be good printing practice for newspapers, which included avoiding monotony. In a few cases the wording of the headline was changed when it was reprinted in lower case, to cast it in a form which the designer considered to be more suitable for presentation in lower case (cf. Figures 1a and f). The number of columns across which the headlines were printed and the positions of the headlines in the newssheets were never altered. The printing and location of the subsidiary headlines were so understandardized (cf. Figures 2a and c) that for the main comparison of upper-case with lower-case headlines in Experiment 1, the subsidiary headlines were deleted by gluing over them blank pieces of newspaper.

The stenciled question sheets contained a list of headlines, with a blank line below each headline where the first word of the first paragraph of news printed under the headline had to be inserted. The order of the headlines in the list was such that the reader had to jump from side to side of the newssheet in going from one question to the next. The large headlines at the top of the newssheet were listed first, followed by the smaller headlines lower down the newssheet. The order was identical for the three printed versions of each of the two newssheets, but since the wording of some of the headlines was slightly altered when they were reprinted, six separate question sheets had to be constructed, so that the wording on a question sheet always corresponded exactly with the wording of the corresponding printed headline.

One set of six question sheets contained the full headlines, main and subsidiary if present. The double headlines were typed in the order in which they were printed, with the subsidiary headline coming first when it was printed as a strap line, but coming second when it was printed as a second deck. There was no indication on the question sheets as to which headline was main and which was subsidiary. The two headlines were typed on separate lines, with a blank line for the answer below the lower headline. Another set of six question sheets contained only the main headlines, omitting the subsidiary. Each of the 12 question sheets was stenciled twice, once typed all in upper-case Olivetti Lexikon 80 elite letters, and once typed with the headlines in the corresponding lower case (capitals only for first letters of headlines and proper names) and underlined, making altogether 24 different stenciled question sheets.

TABLE 1

MEAN NUMBER OF HEADLINES FOUND IN 40 SEC. IN EACH CONDITION

Experiment number	N		Headlines present on		Headlines found in 40 sec. when printed in			Percentage difference	p
	Women	Men	Question sheet	Newssheet	All upper case	Mixed upper and lower case	All lower case		
				Main	6.5		7.1	9	< .01
1	63	9	Main	Main		7.4		9	< .05
			Main	Main + Subsidiary		6.8			
2	24	0	Main	Main + Subsidiary	6.6		7.2	9	< .05
3	12	28	Main	Main + Subsidiary	6.6	6.6		0	*ns*
4	28	12	Main	Main + Subsidiary		6.7	6.5	3	*ns*
5	16	8	Main	Main + Subsidiary	6.5	5.4		17	< .001
6	32	0	Main + Subsidiary	Main + Subsidiary		5.2	4.6	12	< .05
7	27	5	Main + Subsidiary						

Experimental Design

The seven experiments are listed in Table 1. Each experiment compared a single pair of conditions on a separate group of volunteers; nobody performed in more than one experiment. A 2×2 factorial design was always used for two versions of the headlines and the two newssheets. Half the volunteers in each of the four subgroups had question sheets typed in upper-case letters, the other half had question sheets typed in lower-case letters and underlined. The number of volunteers tested was increased either until a reliable difference was found on analysis of variance, or until it was clear that the difference was so small that a large number of volunteers would be required in order to obtain reliability.

Procedure

The experiment was conducted on groups of up to 20 volunteers seated at tables. The level of illumination on the tables ranged from about 40 to 85 ftc. Each volunteer had two sheets of newspaper lying face down on the table and a file containing both the stenciled question sheets and also a practice question sheet and two small practice pieces of newspaper. Usually the practice pieces contained headlines which were samples from the same population as the headlines of the first of the two test newssheets. But in comparing mixed upper- and lower-case headlines with all lower-case headlines, practice pieces with headlines printed in both styles were not available. Instead the first practice piece had all lower-case headlines; the second practice piece had mixed upper- and lower-case headlines. The practice pieces were used to teach the procedure, and immediate knowledge of results was given after each step. Incorrect answers were corrected, and the error was explained to the individual.

The volunteers were told to search the newssheet for headlines in the order in which they were listed on the question sheet. They had to write the first word of news printed under the headline in the blank line provided on the question sheet. Volunteers who had evidently searched the question sheet for headlines in the order in which they located them on the newssheet were discarded, since the aim of the experiment was to compare the speed of searching for headlines on the newssheets, not on the question sheets.

Before each part of the experiment proper the volunteer took a stenciled question sheet from the appropriate compartment of the file, and placed it ready for writing on, on his right if he was right-handed. When everyone was ready all the volunteers turned over their sheets of newspaper (which were on the left if they were right-handed) and started answering the questions. Forty seconds only were allowed for answering the 12 or 14 questions on the question sheet. Only one person ever answered all the questions in that time. The experiment and practices together took about a half hour.

Experimental Subjects

The 264 volunteers were members of a panel maintained at the Applied Psychology Research Unit at Cambridge. Their ages ranged from 21 to 66. The sex distribution of the different groups is shown in Table 1. About half wore reading glasses for the experiment, and a further 9 said that they would have done so had they remembered to bring them. They were paid approximately \$1.05 per hour (i.e., 7 shillings, 6 pence) for their services plus traveling expenses.

RESULTS

The results are given in Table 1. The first experiment used only the main headlines; the

subsidiary headlines were deleted from the newssheets and were not included on the stenciled question sheets. The top row of the table shows that the main headlines printed in lower case were located reliably ($p < .01$) faster than the same headlines printed in capitals. The differences in favor of the lower-case headlines averaged 9%. This is in line with the recommendations made by Paterson and Tinker (1946) from the experiments of their students which have already been referred to. The result could however have been due to the greater size of the lower-case letters used in the headlines (Figure 1f, bottom of pair) compared with the upper-case letters (Figure 1a, top), and to the greater area of paper which they occupied. In this experiment the choice of the lower-case headlines was left to the typographic designer, so that they could be what he considered to be optimal for their particular setting, while the upper-case headlines were those in current use in the newspaper.

The results of the second experiment are given in the second and third rows of Table 1. They show that for the newssheets with mixed upper- and lower-case headlines, the main headlines were located 9% faster when they were the only headlines present on the newssheets, than when the subsidiary headlines were also present ($p < .05$). The presence of the subsidiary headlines apparently distracted the volunteer's attention from the main headlines. Four out of the five subsidiary headlines were strap lines printed in small capitals (three in italic capitals as in Figure 1e) above the corresponding main headline, which was printed in large lower-case letters. The fifth subsidiary headline was printed in small lower-case letters below its main headline.

When the subsidiary headlines were not deleted from the newssheets, the relative advantage of upper- to lower-case headlines depended upon whether the stenciled question sheets gave only the main headlines, or included also the subsidiary headlines. In Experiments 3, 4, and 5, only the main headlines were listed on the question sheets, but both the main and subsidiary headlines were visible on the newssheets. Experiment 3 compared the upper-case headlines with the lower-case. The results were very similar to those of Experiment 1 in which the subsidiary headlines had been deleted from the newssheets: lower case 9% faster ($p < .05$). But here the interpretation is less clear cut, because apart from the upper-case subsidiary strap headline in italics toward the bottom of Figure 2a (which was No. 10 on its question sheet and most people did not reach) the five remaining subsidiary upper-case headlines on the newssheets looked very much like the main headlines below which they were printed. As illustrated in Figures 1a and b, they simply had slightly smaller letters, usually about 3.0 mm. tall compared with the usual 4.0 mm. of the main headlines. Whereas three of the five subsidiary lower-case headlines were in italics (Figure 1f), and all but the one on the left of Figure 3c were considerably smaller than the main headlines, having an x-height of about 1.5 mm., compared with the usual 4.0 mm. of the main headlines. Thus the volunteers with the upper-case headlines may have searched through more headlines on the newssheets in order to locate their targets than did the volunteers with the lower-case headlines.

As in Experiment 3, in Experiments 4 and 5 only the main headlines were listed on the question sheets, while both the main and subsidiary headlines were visible on the newssheets. The table shows that under these conditions there were no reliable differences between the mixed upper- and lower-case headlines and either the all capital headlines or the all lower-case headlines. Clearly it would have been necessary to have tested a lot more volunteers in order to have found reliable differences here, and the interpretations would have been complex.

In Experiments 6 and 7 the stenciled question sheets carried both the main and the subsidiary headlines. The members of each pair were duplicated one below the other in the same relative position as in their newssheet, but it was not possible to tell from the question sheet which was the main and which the subsidiary headline. In the rush and excitement of the experiment it is probable that many volunteers paid particular attention only to the first headline of a pair on the question sheet, and started at once searching for it. This favored the upper-case headlines,

since in the five double headlines at the top of the newssheets the main headlines always came first, as in Figures 1a and b. If the volunteer read only the first headline of a pair on his question sheet, and then searched only the main headlines on the newssheet, he would find the main headline he was searching for. Whereas with three of the four double headlines at the top of the mixed- and lower-case-headline newssheets, the subsidiary strap line came first, as in Figures 1e and f. If the volunteer read only the first headline of these pairs on his question sheet, and then searched only the main headlines on the newssheet, he would be less likely to spot the subsidiary headline which he had read and was searching for. The results of Experiment 6 in the one from the bottom row of Table 1 are compatible with this interpretation. Reliably ($p < .001$) more headlines were located per minute in the newssheets when they were printed in upper case than when they were printed in mixed upper and lower case. The difference in favor of the upper-case headlines was the largest in the table, 17%.

The reliable ($p < .05$) advantage shown in Experiment 7 by the mixed headlines over the lower-case headlines must be due largely to the printing of the subsidiary headlines, since six out of the seven main headlines at the tops of the newssheets were identical in the two cases. The three subsidiary strap headlines which preceded a main headline at the top of a newssheet were printed in 12-point capitals in the mixed-headline newssheets (Figure 1e), whereas they were printed in 12-point lower-case letters in the lower-case headline newssheets (Figure 1f). Presumably when the volunteer had read a subsidiary headline which came first of a pair on his question sheet, and was hunting through the main headlines on the newssheet for it, he was more likely to spot the subsidiary headline if it was printed in small capitals than if it was printed in small lower-case letters. This is in line with Paterson and Tinker's (1946) recommendation that although lower-case letters should normally be used for headlines, capitals can be used occasionally to attract attention.

In addition to the reliable effects of headlines shown in Table 1, there were always reliable effects of order in the analyses of variance, and usually also of newssheets and of individuals. In one of the seven analyses there was a just-reliable ($p < .05$) triple interaction between headlines, order, and newssheets. The lettering of the stenciled question sheets, whether all capitals or lower case (with capitals only for first letters of headlines and proper names) and underlined, reliably affected performance only in Experiment 7, where the lower-case sheets come off better at the .05 level. The lettering of the question sheets did not interact appreciably with the printing of the headlines.

DISCUSSION

The implications of these results are clear:

1. If there is a choice between upper and lower case, headlines will be more discriminable when printed as in this experiment in large bold lower-case letters than when printed in Titling upper-case letters whose height is about the same as the x-height of the lower-case letters.

2. If subsidiary headlines are paired with main headlines set in lower-case letters, they will have a greater attention-getting value when printed in capitals than when printed in lower-case letters of the same point size.

3. But if it is the main headlines which it is desired to emphasize, subsidiary headlines should be regarded as distractors.

REFERENCES

FORBES, T. W., MOSCOWITZ, K., & MORGAN, G. A comparison of lower case and capital letters for highway signs. *Proceedings of the Highway Research Board*, 1950, **30**, 355–373.

MOORE, R. L., & CHRISTIE, A. W. Research on traffic signs. *Engineering for Traffic Conference*, July 1963, 113–122.

PATERSON, D. G., & TINKER, M. A. Readability of newspaper headlines printed in capitals and lower case. *Journal of Applied Psychology*, 1946, **30**, 161–168.

POULTON, E. C. Letter differentiation and rate of comprehension in reading. *Journal of Applied Psychology*, 1965, **49**, 358–362.

TINKER, M. A. *Legibility of print*. Ames: Iowa State University Press, 1963.

(Received October 21, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 426–431

SOME EFFECTS ON BUSINESS GAMING OF PREVIOUS QUASI-T GROUP AFFILIATIONS [1]

SAMUEL D. DEEP, BERNARD M. BASS, AND JAMES A. VAUGHAN

Management Research Center, Graduate School of Business, University of Pittsburgh

93 graduate business students at the University of Pittsburgh were assigned to 9 "companies" to play the Carnegie Tech Management Game. The game was the major portion of a 15-wk. course in integrated decision-making. Men were assigned to "companies" according to whether they had been in the same or different quasi-training (T) groups 15 wk. earlier. "Companies" composed of 2 and 3 subdivided quasi-T groups performed significantly more effectively in the game than "companies" made up of wholly intact quasi-T groups. The latter reported less internal conflict but appeared to be less effective as companies because of overconfidence in each other's dependability.

Much has been written about the effects of the existence of subgroup loyalty in the business organization. A general conclusion of the research in this area is that subgroup loyalty often results in conflict between the overall goals of the organization and the goals of cliques or departmental groups. In this study an attempt was made to create an environment where some business game firms would experience such conflict. Some of the teams were formed with the idea of making them as free as possible of conflicting cliques while others were designed so as to contain either two or three supposed cliques.

It was expected that teams of wholly intact quasi-training (T) groups would achieve a higher level of performance in a business game than teams whose members were drawn from several different quasi-T groups. In particular it was felt that the greatest amount of conflict (and therefore poorest performance) would occur in the teams constructed of men who had previously belonged to different quasi-T groups. On the other hand, "companies" whose members all had been together previously in the same quasi-T group were expected to be at an advantage, because all members had had more experience in working together and tended to know each other a little better. It was thought that such teams would have fewer problems in the communication of ideas and directives and in the coordination of company plans and actions so necessary in the relatively complex Carnegie Tech Management Game,[2] a game which is a reasonably valid simulation of the detergent industry.

METHOD

Procedure

Although the actual experiment involved team performance in a business game over a 15-wk. period, the experience of the participants in the preceding 8 mo. was critical in the formation of teams and thus requires some explanation. Ninety-three graduate business students were formed into nine quasi-T groups of from 9 to 11 members each, in the beginning of a course in behavioral science. (They were identified by letter as Group R, Group S, . . . Group Z.) Each group met twice a week for 15 wk. during the first term, September through December. They are referred to as quasi-T groups because they were trainerless and the group meetings were spread over a considerable amount of time in contrast to a regular intensive 1- or 2-wk. T-group lab with in-group trainers.

In addition to approximately 30 hr. of process analysis (without trainers), each quasi-T group took three midterm examinations together where grades depended on group improvement as well as individual performance. Thus task-relevant, interpersonal interaction was associated with the quasi-T group experience.

At the beginning of the second term each of the nine groups, at the instructor's request, divided itself

[1] This research was supported by Contract Nonr 624(14), Group Psychology Branch. The authors wish to thank John Haas, Susan Wolpert, and Walter McGhee for their assistance in various stages of data collection and analysis. The authors also are indebted to Richard Karpinnen and Ray Bilski for their services as game administrators.

[2] For a complete description of the game see K. J. Cohen, W. R. Dill, A. A. Kuehn, and P. R. Winters, *The Carnegie Tech Management Game.* Homewood, Ill.: Richard D. Irwin, Inc., 1964.

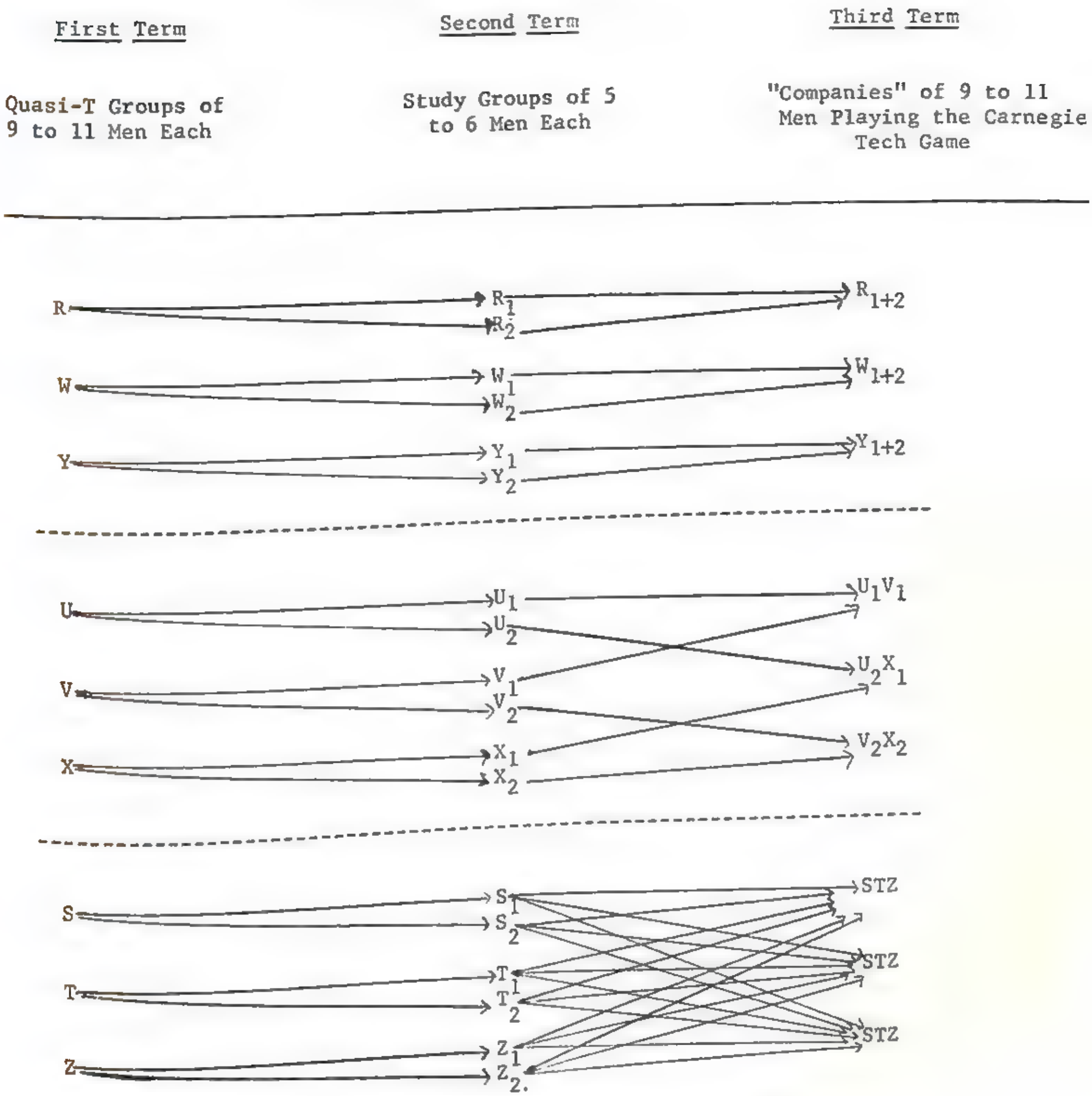


FIG. 1. Plan of separation and mergers.

in half, thus forming 18 study groups for a course in organizational analysis. These study groups met and prepared group papers over the course of the 15-wk. term (January through April).

For the purposes of this study in the third term, it was decided to form nine teams as follows. Three companies were composed of three intact quasi-T groups from the first term. Three other companies were formed from two of the second-term study groups, each of which was from a different quasi-T group. The three remaining companies were formed by bringing together three-thirds of three different quasi-T groups. Figure 1 provides a graphic picture of the above description.

The splitting by self-selection into two study groups for an additional 15 wk. served to increase the possibilities of factionalism under all three conditions. The intact quasi-T groups faced a number of problems: possible guilt feelings about their choice of one subgroup over the other when the T groups had had to divide; unresolved conflicts and resentments over T-group issues which had resulted in the mutual rejections leading to the two clusters which now had been put back together again; and possible unresolved power struggles which were now to be renewed.

The half-and-half teams now had the problem of joining two highly cohesive halves of two different quasi-T groups together who were likely at this time to have different norms, standards, and procedures. Maximum potential for intergroup conflict within the team was likely.

The "three-thirds" teams were the most fractured, had the least experience working together, and thus

TABLE 1

EXPERIMENTAL DESIGN:
INITIAL FINANCIAL POSITIONS OF "COMPANIES" AS REFLECTED IN STOCK PRICES

Industry	$36	$30	$26
1	R	STZ	VX
2	STZ	UX	W
3	UV	Y	STZ

Note.—Letters refer to designations of the ex-quasi-T groups represented intact, in halves, or thirds in the newly formed "companies." R, W, and Y are intact quasi-T groups. U and V, U and X, or V and X made up "half-and-half" teams and quasi-T Groups S, T, and Z made up "three-thirds" teams. Thus, we had three kinds of companies depending on the backgrounds of their merged subgroups.

had more of the usual amount of "warming up" to do.

It was expected that the intact companies who had shared common T-group experiences would have the mechanisms for dealing with the possible guilt or resentment as well as the possible power struggles. It was guessed that the familiarity of all members working together with each other in the past, both in interpersonal analysis and on midterm examinations, would give them a decided advantage over half-and-half teams or three-thirds teams. Initial inertial difficulties in interpersonal interaction were expected among the most fractured three-thirds teams and even greater intergroup difficulties were anticipated for the half-and-half teams. In turn, it was hypothesized that better interpersonal relations would yield better business decision-making performance in the management game. As will be seen, while intact quasi-T groups did report better interpersonal dynamics, their business game performance was considerably less adequate than the performance of competing firms composed of parts of quasi-T groups.

Design

The make-up of the Carnegie Tech game made it necessary to put the nine companies into three industries of three companies each with intercompany competition occurring only within industries. As each of the three competing companies in an industry begins with different pregame business histories and we wished to investigate the effects of the three types of subgroup mergers described above, a Latin-square design was employed, shown in Table 1, so that in each of three industries, A, B, and C, one of three merger plans was combined with one of three initial financial positions as measured by opening market value of the company's stock.

To repeat, competition among teams took place only within each industry. Each company within that industry was given a different past-operating history reflected in the differing initial market values of the stocks of the companies. One company in each industry began the game with the same past financial history. The personnel of one company in each industry began the game with a comparable and real-life interpersonal history.

Objective Measures

Of the many possible performance criteria for judging business firms in competition, the ones used here to compare simulated companies were profits earned, ability to forecast sales, control over production costs, and market value of outstanding capital stock. These were seen as most likely to be sensitive to the differential composition of the companies and most reflective of the differential success of the different companies.

Profits. All teams within an industry began operating their firms with somewhat different past histories of sales, assets, and net worth, but the Latin-square design balanced out these effects (assuming there was no interaction between history, team composition, and industry; that is, no error would be introduced if the effects of pregame history were the same from one industry to the next or in one kind of merger plan compared with another. There was no reason to expect otherwise.). Thus, absolute profits were used as one measure of organizational success, recognizing that maximizing total profit was not necessarily the most salient or the most important goal of all teams at all times. Certain financial ratios, such as earnings per share of common stock, might have been more appropriate, but their values were included for the most part in the determination of another criterion discussed later—stock prices.

Forecasting error. A firm's ability in making effective forecasts was seen as an indication of the adequacy and ease of planning and coordination among its members. Each time that a game-period decision was submitted to the computer, teams also were required to submit a forecast of what they expected sales and profits to be for the coming period. The percentage of error of the actual sales level relative to the forecast was calculated. The forecasts for sales were the ones utilized in this analysis since the profit forecasts were usually very poor for all companies involved.

Avoidable costs. While the costs of poor planning do have an eventual effect on profits and other financial measures, they are of special interest in that they give a direct and more detailed indication of the effectiveness of production coordination taking place. The Carnegie game does a good job of effectively simulating most of the intricacies of timing, coordination, and stabilization needed in the production planning and control aspect of a typical manufacturing concern. In playing the game, decisions must be made about such matters as how much of what raw materials to purchase in order to avoid both supply shortages and excess storage charges, how many workers must be hired or laid off monthly, and how much finished product must be made and supplied to retailers to avoid stockout. To quantify the avoidable costs of poor planning which resulted in runouts, unstable employment, and stock-

outs, an index was calculated to include excess raw material and finished goods storage charges, payment made for unassigned man-hours and unrealized purchase discounts.[3] (Purchase discounts could not be realized when a company did not have the cash available to pay early or on time for raw-material shipments.) Total costs of poor planning were calculated as a percentage of sales since as sales increased so did associated production costs.

Stock prices. The theoretical market value of each firm's outstanding capital stock was computed quarterly (every fourth game month or period). This price was obtained by increasing the earnings per share of the company over its last 4 mo. of operation by a certain multiplier. This multiplier was designed to represent a price-earnings ratio and ranged between 10 and 20. The price-earnings ratio was reviewed each quarter and was raised or lowered (a maximum of 10%) according to the firm's current earnings trend, announcements of dividends, external financing, etc. It was also influenced to some extent by an overall subjective evaluation of the team's performance and interest in the game as estimated by the game administrators in regular meetings with company representatives. The actual stock-value measure used was the percentage of change in stock price at the end of the game over what it was at the start.

The game was played for 30 periods.[4]

Subjective Assessment

A real month after play began, at about the tenth period in game time, each participant completed the following questionnaire:

1. How hard is it for the company as a whole to meet face-to-face?

   1. very hard; 2. fairly hard; 3. neither hard nor easy; 4. fairly easy; 5. very easy

2. How hard is it for you to contact any other member of the company when you need to discuss company business?

   1. very hard; 2. fairly hard; 3. neither hard nor easy; 4. fairly easy; 5. very easy

3. How familiar are you with the other men in your company? To what extent do you know their strengths and weaknesses; and likes and dislikes?

   5. very familiar; 4. fairly familiar; 3. neither familiar nor unfamiliar; 2. fairly unfamiliar; 1. very unfamiliar

4. How easy is it for you to say what you really are thinking to the others in your company?

   1. very hard; 2. fairly hard; 3. neither easy nor hard; 4. fairly easy; 5. very easy

5. How similar are all the men in your company in attitudes, values, goals and abilities?

   5. very similar; 4. fairly similar; 3. neither similar nor dissimilar; 2. fairly dissimilar; 1. very dissimilar

6. If you were asked to assemble a company from the entire class, what are the names of the first five men you would pick?

Company means were calculated for the first five questions. The last question, a measure of cohesiveness, was analyzed according to the differential amount of in-company versus out-company choices made by each member within a company.

RESULTS

Interrelations among Outcomes

Table 2 shows the intercorrelations among the cumulative outcomes for the nine companies over the 30 periods of play.

Objective Outcomes

Improvement in stock prices depended on a variety of company financial factors as well as instructor judgments about overall satisfactoriness of company operations. Nevertheless, it can be seen from the correlation of .94 between profit and stock prices that cumulated profits was a paramount determiner of gain in the price of stock. As might have been expected, these profits and stock prices were adversely affected by poor planning (−.53, −.57), and lack of forecasting accuracy (−.58, −.58).

It also can be seen that the companies who started the game with better stock prices were at an advantage. They showed relatively more profit (.55) and gain in stock prices (.52). They exhibited less poor planning (−.45) but were not superior in forecasting (−.01).

Interpersonal Outcomes

The subjective assessments taken after a month of play tended to meaningfully cluster together also. Cohesiveness, as measured by

[3] Unfortunately, the game did not permit the calculation of the potential existing market in cases where companies failed to meet the demand for their product because of production inadequacies. This opportunity loss should have been incorporated into the cost of poor planning also.

[4] Although the actual game play ended with Period 36, the analysis stops at Period 30. The reason is that Period 31 was the beginning of multiperiod moves (two team decisions due simultaneously) and several bugs appeared in portions of the computer program at this time.

TABLE 2

Product-Moment Correlations among the Objective and Subjective Measures of Company Performance ($N = 9$)

Measure	Profit	Costs of poor planning	Forecast error	Gain in stock price	Starting price	Group contact ease	Individual contact ease	Perceived familiarity	Perceived openness	Perceived similarity	Cohesiveness
Profit		−.53	−.58*	.94**	.55	−.19	−.14	−.03	−.42	−.61*	−.17
Costs of poor planning			.48	−.57	−.45	.30	.07	.11	.66*	.37	.29
Forecast error				−.58*	−.01	.75**	.63*	.59*	.31	.25	.67*
Gain in stock price					.52	−.37	−.10	.05	−.44	−.75**	−.27
Starting price						.24	−.11	.21	−.75**	−.57	.04
Group contact ease							.60*	.46	.26	.39	.75**
Individual contact ease								.58*	.51	.22	.60*
Perceived familiarity									.22	−.04	.82**
Perceived openness										.67*	.44
Perceived similarity											.42
Cohesiveness											

Note.—$N = 9$.
* $p < .05$ with 7 df; $r = .58$.
** $p < .01$ with 7 df; $r = .75$.

in-group and out-group choices, correlated with ease of contact (.75, .60), perceived familiarity, openness, and similarity (.82, .44, .42). Familiarity related to ease of group and individual contact (.46, .58), but openness primarily was associated with ease of individual contact (.51) although the latter was related to ease of group contact (.60). Openness and perceived similarity were positively related (.67).

Objective versus Interpersonal Outcomes

Contrary to expectations, profits and stock prices were higher in those companies which were seen by their members to be divided, that is, not similar in attitudes, values, and goals (−.61, −.75). Perceived openness and ease in saying what one thought hindered, not helped, financial success (−.42, −.44). Costs of poor planning were higher where members felt open (.66); sales forecasting errors were greater in the more cohesive companies (.67) and where group and individual contacts were easier (.75, .63). In short, better interpersonal relations tended to go hand in hand with poorer business decisions.

Experimental Effects

As was expected, the intact quasi-T groups did report the greatest interpersonal satisfaction compared to the halves of two quasi-T groups or the three-thirds of quasi-T groups. As shown in Table 3, contacts were easier, familiarity was greater, more openness was seen, greater similarity of attitudes and purpose was perceived, and more in-group rather than out-group choices were made by the fully reconstituted T groups. Probabilities that the means shown were distributed by chance also are noted in Table 3.

At the same time, Table 3 is consistent with the intercorrelation analysis in that the reverse trend was apparent in the effectiveness of the business decisions made by the three types of companies. The intact quasi-T groups lost 5.37 million dollars instead of showing profits as did the other types. They had the highest costs due to poor planning, exhibited greater sales forecasting error, and yielded the least gain in stock value.

These generally statistically significant results could not be attributed to initial differ-

TABLE 3

DIFFERENCES BETWEEN COMPANIES BALANCED BY INDUSTRY AND INITIAL FINANCIAL POSITIONS ACCORDING TO THE BACKGROUNDS OF THEIR MERGED SUBGROUPS

	Backgrounds of merged subgroups			
	Intact quasi-T groups ($n = 3$)	Halves of two quasi-T groups ($n = 3$)	Thirds of three quasi-T groups ($n = 3$)	All ($N = 9$)
Business				
Profits (in millions)	−5.37	1.97	0.17	−1.08
Costs of poor planning (as % sales)	3.30	0.97	2.73	2.33
Forecast error (%)	56.00	18.30	20.60	31.70**
Gain in stock prices	0.55	2.08	1.47	1.47
Interpersonal				
Group contact ease	3.26	2.41	2.38	2.68**
Individual contact ease	3.76	3.26	2.86	3.29*
Perceived familiarity	4.26	3.75	3.44	3.82**
Perceived openness	4.06	3.78	3.76	3.87
Perceived similarity	3.08	2.66	2.78	2.84
Cohesiveness	2.96	1.81	1.66	2.14**

* $p < .05$.
** $p < .01$.

ences in the ability or intelligence of the participants. Corresponding Latin-square analyses of the companies according to their members' intelligence-test scores (on the Admissions Test for Graduate Schools of Business) and their grade-point averages failed to reveal any differences. In fact, the variances of mean intelligence and grade-point averages of the companies were very near zero.

Game administrators who observed the teams in action felt that the intact quasi-T groups tended to have too many chiefs and not enough Indians. The net result was inability to effectively organize themselves for the real work to be done. One president of an intact quasi-T group complained during the game that his company's consistently poor performance was due to the fact that he knew his teammates too well to be able to control or organize them effectively. He could not needle them for the required amount of work that each had to do if adequate decisions were to be made.

The mergers of men from different quasi-T groups resulted in more structured organizations. Chief executives were more willing to prod subordinates into producing. More important, they were concerned about control and follow-up. A common problem in the companies of men all from the same quasi-T groups was that when a president asked a colleague to complete some analysis, the president assumed too readily that the analysis would be completed on time as requested. There was a reluctance in these companies to initiate and enforce controls that other-type companies used to achieve better decisions. In a sense, there was too much trust and confidence that fellow company members would do their share so that the company would be prepared when deadlines occurred. Among companies of men from different quasi-T groups there was more dependence on management controls and less on familiarity, ease of contact, or perceived similarity of goals and attitudes.

If the results obtained here can be repeated under similar conditions, they suggest that in groups called upon to make many complex decisions under considerable time pressure, the familiarity, cohesiveness, and ease of communications generated by common previous T-group experience may hinder rather than help generate adequate decisions.[5]

[5] It should be emphasized that it would be a mistake to infer from these results that T grouping, per se, contributes to organizational ineffectiveness. Rather, as suggested elsewhere by one of the authors, the T-group experience may or may not be useful for organizational development, depending on what other approaches are followed along with it. For further discussion about this, see B. M. Bass, The Anarchist Movement and the T-Group: Some Possible Lessons for Organizational Development. *Applied Behavioral Science,* in press.

(Received October 31, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 432–435

INFLUENCE OF FAVORABLE AND UNFAVORABLE INFORMATION UPON ASSESSMENT DECISIONS [1]

JAMES W. MILLER AND PATRICIA M. ROWE

University of Waterloo, Ontario

The differential influence of favorable and unfavorable information on assessment decisions was examined. 5 sets of 100 hypothetical stimulus persons were constructed. 32 Ss made assessment decisions to descriptions containing 3 favorable adjectives and 1 unfavorable adjective (3F–1U), 53 Ss to descriptions of the Type 3F–2U and 3F–3U, 30 Ss to descriptions of the Type 2F–3U, and 30 Ss to descriptions of the Type 1F–3U. The results demonstrated that in 4 of the 5 conditions (3F–2U, 3F–3U, 2F–3U, 1F–3U) the unfavorable adjectives were more influential upon the assessment than were the favorable adjectives. The significance of the results for assessment decisions in the employment interview was discussed.

Personnel interviewers and other individuals who have been concerned with making assessment decisions have suggested that unfavorable information about an individual is more influential in making a decision than is favorable information. Yet the literature concerned with the employment interview has only indirectly tested this notion. It was an incidental finding of Crissy and Regan (1951) which suggested that interviewers tend to rely more upon unfavorable evidence about an applicant than they do upon favorable evidence. In addition, there is some evidence (Springbett, 1958) that the employment interview is primarily a search for negative information. That is, the interviewer looks for instances which would lead to rejection of the applicant and if these instances do not appear the applicant will, in all probability, be accepted.

Bolster and Springbett (1961) tested the notion that an early positive decision in the employment situation could be reversed easier than an early negative decision. For those subjects (*S*s) who had an experimentally induced negative decision, it took, on the average, 8.8 items of favorable information to shift the decision from negative to positive. For those *S*s with an experimentally induced positive decision, the shift from positive to negative required only 3.8 items of unfavorable evidence. Thus, their results give support to the notion that unfavorable evidence is more influential upon the interviewer's final decision than favorable evidence.

In the area of person perception, investigators have also been concerned with the differential influence of traits. Asch (1946) developed the concepts of "central" and "peripheral" to refer to traits which have, respectively, great influence and little influence. His problem was to determine how an *S* organizes a series of trait names about a single person into a unified impression; more specifically, he was interested in the effect of changing one trait into its opposite, keeping the other traits the same. The changing of some traits produced marked differences in final impressions while the changing of others produced little change in final impression. Asch concluded that a trait can be either "central" or "peripheral" to the perceiver in its influence on the final impression. Wishner (1960) examined Asch's data and found that there was no way to predict which traits would be central and which would be peripheral without individual measures from each *S* on the associations among traits.

A further reanalysis of Asch's data does, however, indicate that the differential influence of favorable and unfavorable traits upon the final impression can be predicted. The unfavorable characteristics in a description

[1] This article is based on a thesis submitted by the senior author to the faculty of graduate studies of the University of Waterloo, Waterloo, Ontario, in partial fulfillment of the requirements for the degree of Master of Arts. The present research was supported by Grant Number 9401-27 from the Defence Research Board of Canada to Patricia M. Rowe.

served to lower the favorableness of the favorable characteristics far more effectively than the favorable characteristics lessened the unfavorableness of the unfavorable characteristics. Thus, the unfavorable adjectives are likely to be "central" and the favorable adjectives are likely to be "peripheral." A study by Bruner, Shapiro, and Tagiuri (1958) showed that one can also predict the kinds of inferences drawn from trait names. Their results demonstrate that the combination of a favorable and an unfavorable trait will lead to inferences of the same sign (favorable or unfavorable) as the trait that led to the larger number of inferences in isolation. The unfavorable traits generated 50% more negative inferences than favorable traits generated positive inferences. Thus, it appears that unfavorable traits are more influential than favorable traits to an individual when judging another person.

Although the data from experiments concerned with decision making in the employment interview or person perception can be interpreted to show that the unfavorable traits will be more influential to the observer, this interpretation is post hoc and has not been tested systematically. This study deals with a specific type of assessment decision—that of assessing individuals for suitability as roommates—and undertakes a direct and systematic investigation of the differential effects of favorable and unfavorable information upon assessment decisions. It is hypothesized that when an individual makes a decision about another person, on the basis of partial information about that person, the unfavorable evidence will have the greatest influence on the decisions.

Procedure

In the present study, *Ss* were presented with written descriptions of prospective roommates, each of which contained a number of traits. The *Ss* made decisions as to whether or not they would desire each "person" as a roommate. For each trait in a description a measure of favorableness or unfavorableness had been previously obtained.

Selection and scaling of items. Twenty *Ss* at the University of Waterloo rated 299 adjectives from the Gough (1952) Adjective Checklist on a 7-point scale ranging from very favorable to very unfavorable. They were instructed to consider each adjective individually and to make their decisions as if the adjectives were descriptive of a person. Ambiguous adjectives were placed in a separate category.

The mean rating and standard deviation were calculated for each of the adjectives. Ninety-nine adjectives, those receiving one or more ambiguous ratings and those with a standard deviation of greater than 1.0, were then eliminated from the list. This was done to insure that the adjectives used in the descriptions were those about which considerable agreement existed as to the degree of their favorableness or unfavorableness. The remaining 200 adjectives constituted the adjective pool for constructing the descriptions.

Construction of the descriptions. Using these adjectives and their mean ratings, 5 groups of 100 different descriptions of prospective roommates were prepared.

For Group I, 10 sets of 3 favorable (F) and 10 unfavorable (U) adjectives were selected. Within each set the adjectives had approximately the same mean rating. Adjacent sets were separated by a difference in mean rating of approximately .30. To construct the 100 descriptions, each of the 10 sets of favorable adjectives was paired with each of the 10 unfavorable adjectives. Thus, each description in Group I was composed of 4 adjectives, 3 favorable and 1 unfavorable (3F–1U). These 100 descriptions were then randomly ordered to make up Form A of the test materials.

The descriptions in Groups II, III, IV, and V were constructed by adding 1 or 2 unfavorable adjectives to or by eliminating 1 or 2 favorable adjectives from each description in Group I. Thus, each description in Group II was composed of 5 adjectives (3F–2U), each description in Group III was composed of 6 adjectives (3F–3U), each description in Group IV was composed of 5 adjectives (2F–3U), and each description in Group V was composed of 4 adjectives (1F–3U). As with Group I, within each set in Groups II, III, IV, and V the adjectives had almost identical mean ratings and adjacent sets were separated by a difference in mean rating of approximately .30.

Alternate descriptions were selected from Groups II and III and randomly ordered to make up Forms B and C of the test materials. Thus, Form B and Form C consisted of 50 descriptions of the Type 3F–2U and 50 descriptions of the Type 3F–3U. The same description did not appear on both forms.

The descriptions from Group IV (2F-3U) were randomly ordered to make up Form D and the descriptions from Group V (1F–3U) were randomly ordered to make up Form E of the test materials. See Table 1 for a summary of the nature of the test materials.

The rating scale. A 6-point rating scale was used. The task of the *S* was to decide the degree to which he would or would not consider each "person" as a roommate and record his decision from 1 (definitely would consider this person) to 6 (definitely would not consider this person).

Subjects. The *Ss* were 145 undergraduate students in the introductory psychology course at the Uni-

TABLE 1

SUMMARY OF THE TYPES OF TEST MATERIALS

Form	Type of description	No. descriptions
A	3F–1U	100
B	3F–2U	50
	3F–3U	50
C	3F–2U	50
	3F–3U	50
D	2F–3U	100
E	1F–3U	100

versity of Waterloo who received class credit for participating in the experiment.

Method. The test materials were presented to *S*s in group sessions. Thirty-two *S*s responded to Form A of the descriptions, 27 *S*s responded to Form B, 26 *S*s responded to Form C, 30 *S*s responded to Form D, and 30 *S*s responded to Form E. The *S*s were given the forms, asked to read the instructions carefully and respond to all descriptions.

Scoring of responses. In each group of 100 descriptions, each set of favorable and each set of unfavorable adjectives appeared in 10 descriptions. Summing the decisions to the 10 descriptions containing a particular set of adjectives gave the mean decision response (mdr) for all descriptions containing that particular set of adjectives. For example, in descriptions of the Type 3F–1U the set of favorable adjectives *honest, intelligent, original* appeared in 10 descriptions. The mean of the 10 responses by an *S* to these descriptions is the mdr for the set of adjectives *honest, intelligent, original.* This procedure was followed for each set of favorable and each set of unfavorable adjectives for *S*s individually for each of the five types of descriptions.

RESULTS

The care taken in the selection and scaling of the items and in the construction of the descriptions permitted the calculation of a simple measure of the differential influence of adjectives. This measure was obtained by correlating the original rating of the adjectives with the mean decision response (mdr) to those descriptions containing the particular adjectives for each *S*. Since the responses of each *S* were scored to provide an mdr to each of the 10 sets of favorable and each of the 10 sets of unfavorable adjectives, two correlations were obtained from each *S* in each condition. The first, $r_{u.mdr}$, was the correlation between the ratings of the unfavorable adjectives and the mdr's to the descriptions containing those adjectives. The second, $r_{f.mdr}$, was the correlation between the ratings of the favorable adjectives and the mdr's to the descriptions containing those adjectives. A positive correlation indicates that the mdr increases (more descriptions are rated unfavorably) as the adjectives become less favorable or more unfavorable. Any difference, then, between $r_{u.mdr}$ and $r_{f.mdr}$ would reflect a difference in the use to which the favorable and unfavorable adjectives are put. Thus, the class of adjectives having the higher correlation between their rating and the mdr to them are influencing the decision process most. The median and range of $r_{u.mdr}$ and $r_{f.mdr}$ for the five types of descriptions appear in Table 2.

Reaction to favorable and unfavorable adjectives. Differences between the number of *S*s having $r_{u.mdr}$ greater than $r_{f.mdr}$ and the number of *S*s having $r_{u.mdr}$ less than $r_{f.mdr}$ for each type of description were tested by means of χ^2 tests. The summary of these χ^2 tests is shown in Table 3. A greater number of *S*s responding to descriptions of the Type 3F–1U had correlations between mdr and favorable items larger than correlations between mdr and unfavorable items than vice-versa. Thus, for most of the *S*s (68.5%) the favorable adjectives were most influential. By contrast, for a greater number of *S*s responding to Descriptions 3F–2U, 3F–3U, 2F–3U, or 1F–3U the correlations between mdr and items were larger for the unfavorable items than vice-versa. In these latter four groups, therefore, the unfavorable adjectives were more influential than the favorable adjectives. The stronger influence of unfavorable adjectives for most of the *S*s increased steadily as the proportion of unfavorable adjectives increased. Thus, the unfavorable adjectives

TABLE 2

MEDIAN AND RANGE OF $r_{u.mdr}$ AND $r_{f.mdr}$ FOR EACH TYPE OF DESCRIPTION

Items in description	$r_{u.mdr}$		$r_{f.mdr}$	
	Mdn	Range	*Mdn*	Range
3F–1U	.593	−.319 to +.890	.728	+.399 to +.907
3F–2U	.825	+.250 to +.986	.565	−.021 to +.836
3F–3U	.857	+.210 to +.920	.440	−.430 to +.910
2F–3U	.851	+.462 to +.963	.712	+.021 to +.922
1F–3U	.883	+.716 to +.958	.512	−.059 to +.899

TABLE 3

TABLE OF χ^2 TESTS BETWEEN SUBJECTS HAVING $r_{u.mdr} > r_{f.mdr}$ AND $r_{u.mdr} < r_{f.mdr}$ FOR EACH TYPE OF DESCRIPTION

Items in description	Frequency of $r_{u.mdr} > r_{f.mdr}$	Frequency of $r_{u.mdr} < r_{f.mdr}$	χ^2
3F–1U	10	22	4.50*
3F–2U	43	10	20.55**
3F–3U	43	10	20.55**
2F–3U	27	3	19.20**
1F–3U	28	2	21.87**

* $p < .05$.
** $p < .001$.

were most influential for 81% of the *S*s in the 3F–2U and 3F–3U conditions, 90% of the *S*s in the 2F–3U condition, and 93% of the *S*s in the 1F–3U condition.

DISCUSSION

The general purpose of the present investigation was to determine whether certain characteristics of a person are more distinctive than others and, hence, more influential in determining the assessment decisions made of that person by another. The main results are clear. The hypothesis is strongly supported: The unfavorable adjectives carry more weight than the favorable adjectives and exert a disproportionate influence on assessment decisions for most of the *S*s.

This effect appears in four of the five conditions (3F–2U, 3F–3U, 2F–3U, 1F–3U). In these conditions, for a significantly greater number of *S*s, the highest correlations are with unfavorable adjectives. These results indicate that favorable adjectives produce less direct variation in response than do the unfavorable adjectives.

The present results can be generalized to decision making as it occurs in the employment interview. First, this study has extended the earlier results (Bolster & Springbett, 1961) which suggested that the unfavorable evidence in the interview exerts more influence than does the favorable evidence, upon the decision. Their results only showed this to be true, however, when the interviewer had made a preliminary decision and the unfavorable evidence appeared later. In the present study, the undue influence of the unfavorable evidence was also shown to exist when all of the available evidence is present and the interviewer can review it before making his decision.

Second, it is inferred that in basing their decision on the unfavorable adjectives, the *S*s tended to select those "roommates" who were *not bad* and tended to reject those who were *bad*. The same inference can be drawn for those decisions made in the employment interview. The decisions of the interviewer may have consequences not only for the applicant but for the interviewer himself. For instance, the interviewer may receive negative reinforcement for choosing a poor applicant and little positive reinforcement for choosing a good one. If the applicant is largely free from unfavorable characteristics the interviewer can accept him for the position with little fear of negative reinforcement. He thus tends to search for applicants who are *not bad* instead of applicants who are *good*.

In conclusion, this study demonstrates that when an individual must evaluate another individual on the basis of limited information, and when that information contains both favorable and unfavorable evidence, the evaluation will be influenced more by the unfavorable evidence than by the favorable evidence.

REFERENCES

ASCH, S. E. Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 1946, **41**, 258–290.

BOLSTER, B. I., & SPRINGBETT, B. M. The reaction of interviewers to favorable and unfavorable information. *Journal of Applied Psychology*, 1961, **45**, 97–103.

BRUNER, J. S., SHAPIRO, D., & TAGIURI, R. The meaning of traits in isolation and in combination. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior*. Stanford: Stanford University Press, 1958. Pp. 278–288.

CRISSY, W., & REGAN, J. J. Halo in the employment interview. *Journal of Applied Psychology*, 1951, **35**, 338–341.

GOUGH, H. G. *The Adjective Checklist*. Palo Alto, Calif.: Consulting Psychologists Press, 1952.

SPRINGBETT, B. M. Factors affecting the final decision in the employment interview. *Canadian Journal of Psychology*, 1958, **12**, 13–22.

WISHNER, J. Reanalysis of "impressions of personality." *Psychological Review*, 1960, **67**, 96–112.

(Received November 4, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 436–441

COMPARISON OF PERFORMANCE OF SALES TRAINING GRADUATES AND NONGRADUATES

WARD A. HARRIS[1] AND NORMAN L. VINCENT

State Farm Insurance Companies, Bloomington, Illinois

In order to determine if LUTC (a life insurance sales training course) has a favorable effect on sales agents' performance, 88 LUTC graduates were compared with an equal number of nongraduates. The groups were matched man for man on state of residence, age when appointed, length of service, and earnings from the sale of auto, life, and fire insurance in the year LUTC was begun. Performance comparisons were then made for the year LUTC was begun and the year it was completed. While the groups performed about the same during the year the sales training was begun, the trained group produced significantly more 1st-yr. life insurance premiums, on significantly larger life policies, during the year training was completed. In addition to the sales training, 2 other possible reasons for these results are discussed.

Training of life insurance agents is usually spoken of as imparting two distinct kinds of knowledge, technical knowledge about life insurance and sales knowledge about life insurance. Technical knowledge concerns the facts about what insurance is, while sales knowledge deals with the approach to selling it. It is meaningful to consider the two kinds of knowledge separately when evaluating their importance for performance as an agent.

It appears that most investigations of the importance of life insurance knowledge for performance as an agent have dealt exclusively with technical knowledge. Such investigations have attempted to find the relationship between scores on a test of life insurance knowledge and various measures of performance as an agent. The Life Insurance Agency Management Association (1952) reported a slight positive relationship between test score and production, but the relationship did not hold up in a later study (Life Insurance Agency Management Association, 1957). Results of the 1957 study indicated no relationship between test score and average size policy, but the more knowledgeable agents did have better persistency records. Baier and Dugan (1956) reported a slight relationship between test score and production, and in a later study (1957) reported a significant relationship between test score and production, and a positive relationship between test score and persistency. In general, then, the evidence appears consistent in showing positive relationships between technical knowledge and life insurance persistency, but is contradictory concerning the relationship between technical knowledge and production.

The present study was concerned with sales knowledge about life insurance, and an attempt was made to determine if a higher level of sales knowledge about life insurance is associated with better performance as an agent. The study dealt with a particular sales training program, the Life Underwriter Training Council Course (LUTC). This is a 2-yr. sales training program consisting of 25 weekly $2\frac{1}{2}$-hr. classes each year. Emphasis of the course is on development of sales skills. Training time is spent on text books, case histories, field projects, and class discussion.

METHOD

The question was posed simply as: do LUTC graduates perform better as agents than do nongraduates?

Subjects of the study were multiple-line agents of State Farm Insurance Companies, and were located in 25 different states across the country. The performance measures of special interest in the study were:

1. First-year life insurance premiums produced.
2. Average size of life insurance policy produced.
3. Number of life insurance policies produced.
4. First-year and second-year life insurance persistency.[2]

[1] Now at Litton Scientific Support Laboratory, Fort Ord, California.

[2] First-year life insurance persistency is measured as the proportion of life insurance issued in a given

5. Auto insurance production and auto loss ratio.[3]
6. Fire insurance production and homeowners loss ratio.[4]

The procedure followed was to observe the performance of two groups of agents, a group of 88 agents who had graduated from LUTC in May 1964 (LUTC group), and a matched group of 88 agents who were not graduates of LUTC (Control group). Matching was done man for man on the following variables during the year LUTC was started.

1. State of residence.
2. Age when appointed as an agent.
3. Length of service.
4. Total earnings from the sale of auto, life, and fire insurance in year LUTC was begun by the LUTC graduate.

To determine if the groups were essentially alike on the matching variables, tests of significance of the mean difference between the groups were conducted for each variable (see Table 1). The two groups were found to be well matched on all matching variables except length of service. The Control group of nongraduates had, on the average, significantly longer service than the LUTC group.

After ascertaining that the two groups were reasonably well matched, their success on the performance variables was compared. Comparisons were made both for the year the LUTC course was started and the year it was completed. It was expected that both groups would improve their performance during the interval between comparisons because of their increasing experience, but if LUTC has a favorable result, the LUTC group's performance should improve more.

TABLE 1

DIFFERENCES BETWEEN LUTC GROUP AND CONTROL GROUP ON THE MATCHING VARIABLES

Variable	$\bar{X}$ LUTC	$\bar{X}$ Control	$\bar{X}d$	SD	t[b]
State of residence[a]					
Age at appointment (yr.)	31.9	32.0	−0.1	.72	−.14
Length of service (yr.)	5.8	6.5	−0.7	.33	−2.12*
Total earnings	$7837	$8528	−$691	397.66	−1.74

Note.—$N = 88$.
[a] Groups were perfectly matched on state so no significance test was necessary.
[b] Student's t test of the mean difference between groups.
* $p < .05$.

Table 2 shows data and significance test results of comparisons of the performance variables. Significance tests could not be conducted on two of the variables, size and number of life policies, for the beginning year, because of lack of data on individual group members. Rough estimates of the average performance on those variables, however, appear to be quite similar. On those variables for which "ratio scale" measurement was available, t tests of the mean difference between groups were conducted. Wilcoxon matched-pairs signed-ranks tests (Siegel, 1956) were used on those variables for which "interval scale" measurement was available.

Though each of the groups contained 88 agents, complete data were not available for every agent on every variable. Where data were missing for one or both members of a matched pair, the pair was excluded from the analysis. Thus, the N used in the t tests varied from 84 to 88 as can be observed in Table 2. The N used in the Wilcoxon tests also varied from 84 to 88 for the above reasons, and were further reduced by the number of pairs with a difference of zero as required by the test.

RESULTS

During the year LUTC was begun, the average first-year life premiums produced by the two groups were essentially the same. The mean difference of $28 in favor of the Control group was not statistically significant. During the year the LUTC group finished its course of study, 2 yr. later in most cases, the mean difference was significant. Both groups showed an increase in average first-year life premium over the 2-yr. period. While the Control group's increase probably illustrates a normal increase over time, the LUTC group's increase is greater, and this is associated with the fact that they graduated from LUTC. When it is remembered that the LUTC group had a shorter average length of

year that is still in force at the end of that year. Second-year life insurance persistency is measured as the proportion of life insurance issued in a given year that is still in force at the end of the following year.

[3] Auto insurance production is measured in terms of auto career points earned. In essence, 2 auto insurance career points are awarded for each new, reinstated, and added section of automobile insurance.

Auto insurance loss ratio is the ratio of total losses incurred in a given year to total premiums written that year.

[4] Fire insurance production is measured in terms of fire career points earned. Each new and renewal policy issued and credited is associated with points as follows: fire insurance = 1 career point, homeowners = 4 career points, and special multiperil package = 4 career points.

On a continuous homeowners policy, career points are credited at 3-yr. intervals. When the amount of insurance on principal dwelling is $15,000 or more, 2 additional career points are added.

Homeowners insurance loss ratio is the ratio of total losses incurred in a given year to total premiums earned in that year.

TABLE 2

DIFFERENCES BETWEEN LUTC GROUP AND CONTROL GROUP ON THE PERFORMANCE VARIABLES

Variable	N	$\bar{X}$ LUTC	$\bar{X}$ Control	$\bar{X}d$	SD	t[a]
First-year life premium in:						
year LUTC began	84	$1521	$1549	−$28	80.92	−.35
year LUTC finished	86	2387	1991	396	167.54	2.36*
Size of life policy issued in:						
year LUTC began	88	$5696	$5695	—	—	—
year LUTC finished	86	8281	7359	922	378.96	2.43*
No. life policies issued in:						
year LUTC began	88	13.5	13.1	—	—	—
year LUTC finished	86	18.0	16.9	1.1	1.49	.74
Auto career points earned in:						
year LUTC began	85	1213	1245	−32	62.84	−.51
year LUTC finished	88	1242	1257	−15	63.47	−.24
Fire career points earned in:						
year LUTC began	85	227	222	5	16.08	.31
year LUTC finished	88	384	355	29	24.50	1.18
						z[b]
Life persistency ratio (first year) in:						
year LUTC began	60	.92	.91	—	—	−.80
year LUTC finished	65	.93	.94	—	—	−.52
Life persistency ratio (second year) in:						
year LUTC finished	82	.78	.80	—	—	−1.46
Auto insurance loss ratio in:						
year LUTC began	82	.45	.51	—	—	−1.87
year LUTC finished	87	.59	.60	—	—	−.20
Homeowners insurance loss ratio in:						
year LUTC began	82	.50	.57	—	—	−.34
year LUTC finished	86	.75	.68	—	—	−.26

[a] Student's t test of the mean difference between groups.
[b] Wilcoxon's matched-pairs signed-ranks test of the difference between group means.
* $p < .05$.

service, their superior production is especially noteworthy. Generally, it has been shown that length of service and production are positively related. In this case, the LUTC group was able to overcome the handicap of shorter service and outperform the Control group.

The average size of life insurance policies issued in the year LUTC was begun was very similar for the two groups. In the year LUTC was completed, however, the average size policy of the LUTC group was $922 larger than the average for the Control group. Both groups showed an increase in average size policy, but the LUTC group had a larger increase which resulted in a significant mean difference between groups.

The average numbers of life policies for the two groups were nearly identical during the year LUTC was begun. While both groups increased their averages during the completion year, the mean difference between groups still was not significantly different from zero. Therefore, there is no evidence that graduation from LUTC is associated with a greater than normal increase in average number of life policies issued.

First-year life insurance persistency comparisons are also shown in Table 2. The difference between groups during the year LUTC was begun was not significant. For the year LUTC was completed, both groups increased their first-year life persistency, but neither group did significantly better than the other. From these data, it appears that graduation from LUTC cannot be associated with significantly better first-year life persistency.

It was thought that a greater difference between groups might show up if one looked at second-year persistency.

As can be seen in Table 2, both groups had considerably lower second-year persistency than they did first-year persistency. Both groups, however, suffered about the same decrease. The difference between the two groups' second-year persistencies is not statistically significant. It might be speculated that life insurance persistency of the two groups will diverge as more time passes. Perhaps third-, fourth-, or fifth-year persistency will show a significant difference. The evidence available now does not appear to support such speculation. Instead, it points to chance fluctuation as causing the only difference in persistency between the two groups.

Let us now look at auto insurance production and auto insurance loss ratio. As noted previously, auto production is measured in terms of auto career points earned. It can be seen in Table 2 that during the year LUTC was begun, the Control group earned, on the average, 32 more auto points than did the LUTC group. During the year LUTC was completed, the Control group's lead was only 15 points on the average. Before applauding the faster increase of the LUTC group, however, it should be noted that the mean differences between the two groups, both when LUTC was started and when it was completed, are not statistically significant. Thus, a conclusive statement of greater auto-production growth for the LUTC group is precluded.

The situation is similar for auto loss ratio. The Control group showed, on the average, a larger loss ratio the year LUTC was started, but during the year of completion the average loss ratios were nearly identical. Again, the only strong conclusion that can be drawn about the two groups is that their difference in auto loss ratios was "due to chance," both when LUTC was begun and when it was completed. Thus, no change can definitely be associated with LUTC.

Comparisons of average fire insurance production, indicated as fire career points earned, are shown in Table 2. In the year LUTC was begun, the difference between average fire points for the two groups is very small and certainly not significant. During the year LUTC was completed both groups show an increase in fire points, but the mean difference between the two groups is still not statistically significant. So, although the LUTC group shows a slightly larger increase in points earned, it is probably a chance occurrence.

Looking at homeowners loss ratio, one sees that the LUTC group had a slightly lower loss ratio when the LUTC training was begun. When training was completed, however, the LUTC group had a slightly higher loss ratio. This appears to be chance fluctuation since neither of the differences is statistically significant. The inference that must be drawn from these data is that LUTC has no significant association with homeowners loss ratio.

DISCUSSION

It can be concluded on the basis of these comparisons that the LUTC group performed better on two important aspects of life insurance production. Caution should be used, however, in concluding that better performance was necessarily a direct result of LUTC participation. It may be that the kind of agents who produce better are also the kind who choose to attend LUTC. There may be a factor of "interest in life insurance" which is basic to both attendance of LUTC and production of life insurance.

The possibility of an "interest factor" is strengthened by the fact that the LUTC group had the same average level of production as did the Control group during the year LUTC was begun, but had a significantly shorter average length of service. For the LUTC group to reach the same level of life production in a shorter period of time, there probably was some factor such as "interest" spurring them on. If an "interest factor" is operating, the LUTC group might have performed better than the Control group even if they had not attended LUTC.

However, if higher "interest in life insurance" is inferred from the LUTC group's faster increase in life production up to the year LUTC was begun, the same reasoning may apply to auto and fire production. The

TABLE 3

INTERCORRELATIONS OF CAREER POINTS EARNED IN 1 YEAR

	Life	Fire
Auto	.60*	.68*
Life	—	.45*

Note.—One career point is awarded for each $2.50 of paid first-yr. life premium (excluding Preliminary Term premium). Awarding of auto and fire points is explained in previous footnotes. $N = 358$.

* $p < .01$.

LUTC group required less time than the Control group to reach the same level of auto and fire production, and, therefore, must have higher "interest in auto and fire insurance." It is quite possible, then, that the LUTC group did have higher "interest in life insurance" than the Control group, but apparently this interest extended to auto and fire insurance as well.

Assuming that the interest level of the LUTC group toward auto and fire insurance as well as life insurance is higher than that of the Control group, one would expect the LUTC group to produce better than the Control group in all three lines. Previous analyses have shown a strong relationship among production levels of the three lines indicating that what influences production of one line also influences production of others (see Table 3). The fact that the LUTC group was significantly higher only in life production indicates that something in addition to interest is probably influencing life production; it is probably the LUTC sales training course.

Another possible reason why the Control group's average life production was significantly less than that of the LUTC group, during the year LUTC was completed, is that the Control group's production may simply be regressing toward a lower population average. It can be assumed that the two groups of agents were samples from two different populations of agents, one a population of all agents who will graduate from LUTC (LUTC population), and the other a population of agents who will not graduate from LUTC (Control population). The average life production of the LUTC population might be higher than the Control population's average both during the year LUTC was begun as well as during the year it was completed. The reason the Control group's average life production was similar to that of the LUTC group's could have been only because a group whose life production matched that of the LUTC group was purposely selected, and the Control group's average could have been considerably higher than the Control population's average. Over a period of time, however, the Control group's average would tend to regress back toward the Control population's average, and by the time LUTC was completed the regression may have been enough for a significant difference between the groups' average to appear.

No figures are available on the population averages; hence, the possibility of the regression phenomenon cannot be refuted. If the significant difference in life production can be explained by regression toward the mean, it is relevant to ask why the same phenomenon does not affect auto and fire production. For both of these, the Control group had average production very similar to that of the LUTC group during the year LUTC was begun, and average production was still quite close to that of the LUTC group even after time had elapsed. Possibly for auto and fire insurance the Control population's average production was never very divergent from the LUTC population's average, and no great amount of regression could take place. Another possibility could be that the rate of regression for auto and fire production is slower than it is for life production and more time would allow significant differences to appear for them also.

What the exact relation between LUTC graduation and performance as an insurance agent is, cannot be ascertained on the basis of these comparisons alone. Nevertheless, LUTC graduation tends to be associated with agents who will generate more first-year life insurance premium and produce larger life insurance policies. If LUTC does no more than help identify agents who will produce high premium and large policies in the future, it will serve a worthwhile function.

CONCLUSIONS

1. LUTC graduates tend to produce more first-year life premium than do nongraduates.

2. LUTC graduates tend to produce larger life policies than do nongraduates.

3. LUTC graduates and nongraduates tend to produce about the same average number of life policies in a year.

4. LUTC graduates and nongraduates appear to produce life insurance business having about the same level of 1-yr. and 2-yr. persistency.

5. LUTC graduates and nongraduates tend to produce about the same amount of auto insurance and have about the same auto loss ratios.

6. LUTC graduates and nongraduates tend to produce about the same amount of fire insurance and to have similar homeowners loss ratios.

REFERENCES

BAIER, D. E., & DUGAN, R. D. Tests and performance in a sales organization. *Personnel Psychology,* 1956, 9, 17–26.

BAIER, D. E., & DUGAN, R. D. Factors in sales success. *Journal of Applied Psychology,* 1957, 41, 37–40.

LIFE INSURANCE AGENCY MANAGEMENT ASSOCIATION. Insurance knowledge: A report on the Information Index. Research Report 1952–5, File No. 450.

LIFE INSURANCE AGENCY MANAGEMENT ASSOCIATION. Insurance knowledge and performance. Research Report 1957–4, File No. 450.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

(Received November 7, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 442–443

SCALING ASSUMPTIONS UNDERLYING WEIGHTING IN JOB-CLASSIFICATION SYSTEMS

FRANK J. LANDY
Bowling Green State University

AND

ARTHUR J. ELBERT
Parsons College

The present study applied Thurstone's Law of Comparative Judgment to a job classification of hourly employees in a food-processing plant. Supervisors rated 11 job elements on importance for overall production using a paired-comparison format. The 11 elements were scaled using 3 different methods: Case V solution, Case III solution, and the Composite-Standard method. Reversals in element positions were found as a function of the scaling method used. The Composite-Standard method appeared to be the least appropriate of the 3 while the Case III solution seemed most applicable, taking the inequality of element dispersions into account. A classification inequity might have resulted had the Case V or Composite-Standard values been used as element weights.

The present study concerned itself with applying Thurstone's Law of Comparative Judgment to a job classification of hourly employees in a food-processing plant. The construction of a point system necessitated the determination of appropriate weights for 11 predetermined job elements. The following is a list of these elements and a brief description of each: (1) Experience—amount of experience required to perform the job duties satisfactorily; (2) Education—amount of formal school training; (3) Responsibility for equipment, tools, product, and materials—job responsibility for preventing damage to equipment, tools, raw product, or partly finished or completed products in the performance of the job; (4) Responsibility for the work of others—responsibility for assisting, instructing, or directing others in the performance of their work; (5) Resourcefulness—independent judgment, making of decisions, or amount of planning which the job requires; (6) Monotony—sameness or lack of variety of the work which might create a feeling of monotony; (7) Pressure of work—the pace at which the employee must work; (8) Physical effort—amount and continuity of physical effort required on the job; (9) Surroundings—conditions under which the job must be performed; (10) Hazards—health and accident hazards connected with the job; (11) Concentration—the degree of concentration required, either to minute job details or many tasks of the job.

TABLE 1

SCALE VALUES OF JOB ELEMENTS USING DIFFERENT METHODS

Job element	Case V	Composite-Standard	Case III
Monotony	.59	1.18	1.60
Surroundings	2.45	2.93	3.00
Hazards	3.12	3.44	2.68
Pressure	3.60	3.69	3.97
Education	4.52	4.64	4.61
Effort	4.75	4.79	4.96
Concentration	5.70	5.62	5.48
Experience	6.64	6.47	6.54
Resourcefulness	7.71	7.21	6.67
Responsibility for equipment	8.23	7.57	7.07
Responsibility for others	7.79	7.62	8.53

Note.—Scale values transformed to $\bar{X} = 5$, $SD = 2$.

The elements were presented to 29 supervisors in a paired-comparison format yielding 55 pairs. They were instructed to choose the element of each pair which they considered most relevant for overall plant production.

The results were analyzed assuming Case V (Guilford, 1954, p. 156). The scale values are presented in the first column of Table 1. A chi-square test of internal consistency (Mosteller, 1951) yielded a significant value ($\chi^2 = 62.73$, $p < .05$, $df = 45$), therefore the scaling was repeated applying a Case III solution. A test of internal consistency of the Case III solution yielded a nonsignificant value ($\chi^2 = 33.42$, $p < .30$, $df = 35$). The

scale values are presented in the third column of Table 1. Estimates of standard deviations of element judgments indicated a wide range of dispersions, the largest standard deviation being about five times the smallest. As a function of nonagreement on the "Surroundings" element, a reversal appeared between the positions of "Surroundings" and "Hazards" in the Case III solution. In addition, when the inequality of dispersions was taken into account, the reversal between "Responsibility for equipment" and "Responsibility for the work of others," found in the Case V solution, disappeared.

The results were also analyzed using the Composite-Standard method (Guilford, 1928). The scale values are presented in the second column of Table 1. This method seemed to be the least appropriate of the three. Neither of the two previously mentioned reversals appeared nor was there an opportunity to justify the assumptions of a Case V solution. It was concluded that the Case III solution provided scale values most appropriate for assigning weights to job elements.

The most important consideration to take into account would seem to be the consequent role of the scale values. In the present context, an inequity in the job-classification system might have resulted if either the Case V or Composite-Standard values had been used as element weights.

REFERENCES

Guilford, J. P. The method of paired comparisons as a psychometric method. *Psychological Review*, 1928, 35, 494–506.

Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

Mosteller, F. Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika*, 1951, 16, 207–218.

(Received December 7, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 5, 444–448

COPYING ALPHA AND NUMERIC CODES BY HAND: AN EXPERIMENTAL STUDY

R. CONRAD AND A. J. HULL

Medical Research Council, Applied Psychology Research Unit, Cambridge, England

Different groups of housewives copied alphabetic and numeric codes by hand under 5 different conditions of copying. Factors substantially affecting speed or accuracy of copying or both are: code length, distance between original code and copied code, whether the code was alphabetic or numeric, and, for 12-digit codes, grouping of digits. Detailed analysis of copying errors points to a marked short-term memory component, and this conclusion is supported by the similarity between the present results and those reported from conventional short-term memory studies.

The increasing use of computers, automatic reading devices, and so on, is forcing a revision of the way individual items of multi-item systems are identified. For some time to come it seems as if these names will take the form of alphanumeric codes. Copying a code by hand onto a form or envelope is likely to become an activity carried out by an increasing number of people. Although as a rule any one person may copy a code in this way relatively rarely, the total number of such code copyings in large national systems like social security, pensions, Giro,[1] savings account, etc., easily runs into millions daily in, for example, the United States or the United Kingdom. Furthermore the people required to do this copying are not a few selected and trained professionals, but the general public in a very wide sense.

In this simple transaction between a man and a complex system, speed of code copying is not necessarily an issue. But accuracy always is. Even with automatic checks built into the code so that erroneous codes are detected with a high probability, a cost is inevitably involved in ascertaining what the code should have been, particularly when, as is common, the "customer" is no longer present when the check is made. Possibly because the easy assumption is that copying letters or digits is too highly practiced a skill to carry an error risk, little if any research data have been published concerned with factors relevant to error rates in alphanumeric code copying by hand.

Klemmer and Lockhead (1962) report error data for digit copying by professional key punchers, and Braunstein and Anderson (1959) examined errors when digits are read aloud. Though related, these studies are not central to the problem. Neither is that by Conrad (1962) which is concerned with the question of whereabout in an alphanumeric code to locate numerals when the code has to be remembered. The present study therefore looks at the problem more directly and compares error rates and speed when codes are presented to the "customer" in different ways. In varying degrees the following factors were investigated: distance between source code and copy, length, configurative grouping within a code, and whether the code was all alpha or all numeric. But in the course of the study some general results emerged which might illuminate the more fundamental psychological processes involved in this simple, widely exercised, and economically important skill.

METHOD

The basic test procedure was simply to provide *Ss* with a booklet of foolscap sheets (8 in. wide) containing one or two columns of random digit or letter sequences, and to instruct them to copy the sequences (codes) as quickly as possible in indicated spaces. Emphasis was on speed for no better reason than to try to generate more useful error data, and checking was discouraged. Tests were for 15 min., and all *Ss* were paid housewives. No *S* took part in more than one test. Test sheets were very clearly duplicated from excellent stencils using pica type (letters in upper case).

[1] A system used in a number of European countries for noncash money transactions using post office facilities instead of banks.

Tests

1. Codes were displayed in a single column at the left-hand edge of the sheet with substantial vertical spacing. A dotted line from each code crossed the sheet, and Ss copied each code to the right of a line drawn vertically 2½ in. from the right-hand edge of the sheet. Codes were 3, 6, 9, or 12 digits in equal numbers but random order. Digits were grouped in threes by a space, for example, 496 261 173. Forty-one Ss were used.

2. This test was identical to Test 1, except that there were no spaces between digits, for example, 496261173. Forty Ss were used.

3. Codes were displayed with no spaces between digits in two columns with substantial vertical spacing. They were 3, 6, 9, or 12 digits long in equal numbers and random order. The Ss copied each code immediately beneath each displayed code. Forty-nine Ss were used.

4. This test was similar in form to Test 3, that is, codes ungrouped and copied immediately below. But the codes were alphabetic, drawn from the 10-letter vocabulary: BCFMNPSTVX. The main reason why this particular vocabulary was used was simply because, having been used in other experiments, some data were available on likely intravocabulary shape and sound confusions. All sequences contained eight letters drawn randomly from the vocabulary. Forty-four Ss were used.

5. The same eight-letter sequences used in Test 4, but with the format of Test 2, that is, ungrouped and copied across the sheet. Forty-three Ss were used.

Summarizing these conditions for quick reference:

1. Digits grouped far.
2. Digits ungrouped far.
3. Digits ungrouped near.
4. Letters ungrouped near.
5. Letters ungrouped far.

The Ss' protocols were scored for speed and accuracy independently by two scorers who had no idea of the purpose of the tests. Scoring discrepancies were then checked by E. The result was a set of highly reliable results minimally contaminated by scorer error or bias.

RESULTS

The results to be presented relate to five independent tests which in some cases were set up to meet a practical requirement. Because of certain confoundings, then, comprehensive intertest comparison is pointless. The tests though did use large separate groups of similar Ss, and where parameters are not confounded performance differences can be usefully measured.

Speed and accuracy of copying in the five tests are shown in Table 1. Because codes varied in length both within and between some tests, the unitary measure of characters per S per minute is used for speed. For accuracy, the more practically useful criterion of whether or not there was an error in the code was used, that is, percentage of wrong codes. Because where it was varied, code length was scrambled within a test, it is not possible to present speed data as a function of length. But later on, length against accuracy is shown.

Table 2 shows the differences between certain tests both for speed and for errors. The effect of grouping digits seems to have no effect on rate of copying measured over all code lengths. If there were a Rate × Length interaction it is unfortunately lost in the experimental design. Proximity though turns out to be important. We can also usefully compare the copying rates of digits and letters according to whether copying is near or far (only tests for ungrouped codes are available). These differences are also highly significant. It is probable that some of the interactions are statistically significant, but because they

TABLE 1

SPEED AND ACCURACY OF COPYING CODES UNDER VARIOUS CONDITIONS

Test condition	Characters copied per S/min	% wrong codes	No. wrong codes
1. Digits grouped far	92.7	1.42	108
2. Digits ungrouped far	90.7	1.75	127
3. Digits ungrouped near	101.5	1.11	90
4. Letters ungrouped near	85.3	2.63	185
5. Letters ungrouped far	69.1	3.15	175

TABLE 2

DIFFERENCES IN SPEED AND ACCURACY OF COPYING CODES UNDER SPECIFIED CONDITIONS

Conditions compared	Speed t	Accuracy χ^2
Digits grouped versus ungrouped	1.0	2.59
Digits near versus digits far	3.05*	24.11**
Letters near versus letters far	5.06**	3.02
Digits near versus letters near	5.46**	74.18**
Digits far versus letters far	5.68**	26.68**

* $p < .01$.
** $p < .001$.

TABLE 3

EFFECT OF CODE LENGTH ON COPYING ACCURACY

Test condition	% wrong codes			
	Code length (no. digits)			
	3	6	9	12
1. Digits grouped far	.47	1.16	1.26	2.79
2. Digits ungrouped far	.33	.99	1.49	4.19
3. Digits ungrouped near	.56	.56	1.29	1.21

would be in the expected direction, and are therefore of little practical value, it has not seemed worthwhile to make the statistical tests.

In the case of accuracy, of the five comparisons made, in three of them significant speed differences are associated with significant error differences in the same performance direction. Indeed looking at the speed and accuracy data in Table 1, it is evident that the relationship is monotonic; that is, a difficult code is copied both slowly and inaccurately. In the case of near versus far letters, the error difference is in the expected direction, but does not reach the .05 level on a two-tailed test (though it would on a one-tailed test).

For three of the five conditions, accuracy can be shown as a function of code length (Table 3). These data throw some light on the apparent absence of either speed or accuracy difference between grouped and ungrouped codes (Tests 1 and 2). For the 12-digit codes there is a significantly greater proportion of errors without grouping ($p <$ .02), and this difference is masked in the overall lengths measure by the absence of effect for the other lengths. Similarly, when ungrouped digits near and far are examined length by length (Tests 2 and 3), it turns out that the significantly different overall effect reported is due largely to the 12-digit code; the differences for Lengths 3-, 6-, and 9-digits do not reach the .05 significance level.

Codes were incorrectly copied by four criteria—the only ones to emerge from an examination of the raw data: a single character substituted, added, or omitted; a pair of adjacent characters transposed. Whether or not in operational conditions a code can tolerate certain types of error more than others clearly depends on the particular usage. Table 4 presents, for each condition, the proportion of errors falling into the four categories of transposition, substitution, addition, and omission (plus or minus one character; no others were found). Once one has taken into account the small number of errors represented, the only striking difference between conditions is in the number of incomplete codes in Test 2. Furthermore most of these are due to the 12-digit code, where it seems likely that *S*s copying in chunks lose their place. The fact that this phenomenon is largely associated with 12-digit codes may explain why it does not markedly occur with the letter codes which all have eight characters.

Again discounting small-sample errors, it is evident that most errors are due to copying a single character incorrectly. Nevertheless the proportion of transposition errors must be far greater than would occur by chance. In fact on every occasion that two adjacent characters were wrong, they were transposed.

The two conditions using letter codes provided most data, and the location of errors in the 8-letter sequence was noted. To create useful sizes of sample, these were classified as occurring in the first or second half of the code (Locations 1–4 and 5–8). For near copying the proportions were, respectively, 51.7% and 48.3%. This difference is not significant. For far copying the values were 42.9% and 57.1%. This difference is significant at the .02 level.

In case in Table 2 the error comparison of 8-character letter codes with digit codes using the overall-lengths error score be regarded as dubious, it can be stated that the significance

TABLE 4

ANALYSIS OF COPYING ERRORS

Test condition	% errors				N
	Transposition	Substitution	Plus 1	Minus 1	
1. Digits grouped far	24.1	66.7	2.8	6.4	141
2. Digits ungrouped far	5.8	33.1	7.2	53.9	139
3. Digits ungrouped near	15.5	77.7	1.9	4.9	103
4. Letters ungrouped near	5.0	86.9	2.0	6.1	199
5. Letters ungrouped far	4.3	80.8	6.9	8.0	188

levels are unchanged when letter codes are compared with 9-digit codes, which if anything would tend to reduce differences. The point does arise, of course, that the unusual letter vocabulary used biased the letter codes detrimentally. Superficially this is a reasonable comment. But it is not easy to see on closer consideration why it should in fact make a difference. In adult Ss all letters of the alphabet are highly familiar, and the particular subset is within 5% of the mean letter frequency of the alphabet as a whole as it occurs in written or spoken English (Baddeley, Conrad, & Thomson, 1960). Furthermore there is no question that Ss need to remember the vocabulary—the copied code is always visible. Nevertheless it is probable that 10-letter subsets could be found from which more efficiently copied codes could be derived. From the magnitude of the presently observed effects, it seems doubtful whether any 10-letter subset would be superior in this situation to digits. But the point does remain open, as does the effect of using other sizes of vocabulary.

DISCUSSION

The more detailed analysis of the kinds of errors made when copying codes points to a substantial short-term memory component. First, the fact that there are transposition errors at all over and above the very few that might occur by chance, suggests forgetting rather than simple failure of mechanical reproduction. Second, the significant increase in errors in the second half of letter codes copied far, is typical of immediate recall data when sequences are initially silently read (Conrad, 1965, 1966; Murray, 1966). Although there is no internal evidence in the substitution data to point particularly either to copying or to memory as the primary error source, it is likely that the evidence for forgetting is masked by the effects of "copying" errors. With an error rate of only a few wrong characters per thousand, an even more detailed error analysis was impractical.

This suggestion that even in the simplest copying task memory is important is indirectly supported elsewhere in the results. Once *S* is obliged to make visual shifts of some duration, perceived material must be briefly held in a memory store; related to this is the likelihood that the greater the distance between the original and the copy the less likely would be running character-by-character monitoring. Thus near-copying is more efficient than far-copying, and, in the letter case, the near-copying errors are not distributed in the sequence like memory errors. Finally the effect of grouping is similar to that reported in analogous recall situations (Conrad & Hille, 1957; Thorpe & Rowland, 1965), and so also is the digit-letter comparison (Crannel & Parrish, 1957; Mackworth, 1964).

Two kinds of practical conclusions are suggested. The first is simply that since ad hoc comparisons of different code structures in a copying situation are costly of *S* time and laborious, relations between codes may more economically be derived from immediate recall tests which will yield far more error data per *S*.

The second concerns the rules in addition to whatever system constraints are necessary for designing operational codes which will be handled, and particularly copied, by people. Clearly, length of code is important especially if there is likely to be a distance even merely of a few inches between the original and the copy. A problem here is that codes are "artificially" lengthened by the addition of check digits when the codes are part of a computerized system. Lengthening codes in this way permits automatic error detection, but not correction. Optimizing is needed as between fewer errors but undetected by the system, and more errors made but detected and requiring correction. Grouping—not unexpectedly—helps.

Perhaps the most useful rule where it is applicable refers to proximity. There seem to be substantial advantages to copying a code immediately beneath the original. In practical terms this becomes a problem of designing forms, dockets, etc., in such a way that this is possible.

On the evidence presented here, digit codes are preferable to letter codes, a conclusion which must be thickly surrounded by qualifications. There are many obvious advantages for numeric codes in computerized systems,

particularly if a decimal vocabulary is a priority. Where larger alphabetic vocabularies are practicable, utilizing language habits (Baddeley, 1964; Underwood & Schulz, 1960) could be advantageous in suitable cases.

One final point here. If code-copying errors resemble code-memorizing errors to the extent of following the commonly reported serial position effect, then the location in the code of particular characters becomes important. That is, the most logically important characters must be located in the psychologically least vulnerable positions (Conrad, 1962).

REFERENCES

BADDELEY, A. D. Immediate memory and the "perception" of letter sequences. *Quarterly Journal of Experimental Psychology*, 1964, **16**, 364–367.

BADDELEY, A. D., CONRAD, R., & THOMSON, W. E. Letter structure of the English language. *Nature*, 1960, **186**, 414–416.

BRAUNSTEIN, M., & ANDERSON, N. S. A comparison of reading digits aloud and keypunching. *IBM Research Memorandum*, 1959, RC-185.

CONRAD, R. The location of figures in alpha-numeric codes. *Ergonomics*, 1962, **5**, 403–406.

CONRAD, R. Order errors in immediate recall of sequences. *Journal of Verbal Learning and Verbal Behavior*, 1965, **4**, 161–169.

CONRAD, R. A short-term memory factor in the design of data entry keyboards: An interface between short-term memory and S-R compatibility. *Journal of Applied Psychology*, 1966, **50**, 353–356.

CONRAD, R., & HILLE, B. A. Memory for long telephone numbers. *Post Office Telecommunications Journal*, 1957, **10**, 37–39.

CRANNEL, C. W., & PARRISH, J. M. A comparison of immediate memory span for digits, letters, and words. *Journal of Psychology*, 1957, **44**, 319–327.

KLEMMER, E. T., & LOCKHEAD, G. R. Productivity and errors in two keying tasks: A field study. *Journal of Applied Psychology*, 1962, **46**, 401–408.

MACKWORTH, J. F. Auditory short-term memory. *Canadian Journal of Psychology*, 1964, **18**, 292–303.

MURRAY, D. J. Vocalization-at-presentation and immediate recall, with varying recall methods. *Quarterly Journal of Experimental Psychology*, 1966, **18**, 9–18.

THORPE, C. E., & ROWLAND, G. E. The effect of "natural" grouping of numerals on short term memory. *Human Factors*, 1965, **7**, 38–44.

UNDERWOOD, B. J., & SCHULZ, R. W. *Meaningfulness and verbal learning*. New York: Lippincott, 1960.

(Received January 16, 1967)

Manuscripts Accepted for Publication in the *Journal of Applied Psychology*

Automobile Accidents and Driver Reaction Pattern: Paul Babarik*: Human Sciences Research, Inc., Westgate Research Park, McLean, Virginia 22101.

Interface between Personnel and Organizational Psychology: Bernard M. Bass*: 401 Bruce Hall, Graduate School of Business, University of Pittsburgh, Pittsburgh, Pennsylvania 15213.

Particularism and Organizational Position: A Cross-Cultural Analysis: Louis A. Zurcher*: The Menninger Foundation, Box 829, Topeka, Kansas 66601.

Differential Effect of Noise on Tasks of Varying Complexity: David H. Boggs and J. Richard Simon*: Department of Psychology, University of Iowa, Iowa City, Iowa 52240.

* Asterisk indicates author for whom the address is supplied.

journal of applied psychology

Vol. 51, No. 6 December 1967

COMPARATIVE MANAGERIAL ROLE PERCEPTIONS IN MILITARY AND BUSINESS HIERARCHIES

VANCE F. MITCHELL and LYMAN W. PORTER

University of California, Berkeley

This study was concerned with comparison of the managerial and administrative role requirements perceived by military officers and civilian managers. A questionnaire provided data from 703 commissioned officers and 594 noncommissioned personnel serving in an overseas Air Force Command. The findings showed that inner-directed traits are regarded as more important for job success than other-directed, as among civilian managers. However, the pattern of the change among levels of the military hierarchy and the primary role requirements perceived differed considerably from those exhibited by civilian managers. Also, commissioned officers' perceptions of the behavior necessary for job success differed greatly from those of noncommissioned personnel. Results from the noncommissioned officers show clear dominance of other-directed behavior.

In a comparative study, Porter and Mitchell (1967) have shown that military officers (in an overseas Air Force Command) experience somewhat greater frustration and dissatisfaction in their job assignments than civilian managers of equivalent level. The question next arises as to whether these two groups are likely to perceive different managerial and administrative role requirements in their respective types of organizations. More specifically, do military officers tend to place relatively greater emphasis on inner-directed behavior as a prerequisite for job success when compared to managers in business firms? Equally important is the question of whether there are changes in such role perceptions within the military hierarchy (both commissioned and noncommissioned officers) that can be associated with rank, as is the case for business organizations (Porter & Henry, 1964).

Pertinent to the above questions are the speculations of several social scientists concerning the existence of conflicting role requirements in the modern military organization. For example, Janowitz (1960) notes that military hierarchies are commonly viewed as

> rigidly stratified and authoritarian because of the necessities of command and the possibilities of war. . . . Moreover, because military routines tend to become highly standardized, it is assumed that promotion is in good measure linked to compliance with existing procedures [p. 8].

He advances the hypothesis, however, that

> the organizational revolution which pervades contemporary society and which implies management by means of persuasion, explanation and expertise, is also to be found in the military [p. 8].

Lang (1965) also has argued to this effect and cites several studies illustrating the existence within military organizations of both a variety of leadership patterns and a degree of conflict between more permissive, interactive management and the traditional, hierarchical structure of the organization. The present study was designed to provide some evidence with respect to these discussions of the military as a particular type of social organization.

Method

Questionnaire

The data for this study were collected by means of the same questionnaire, slightly modified for applicability to military respondents, which was previously used in studies of managers in business firms

(Porter & Henry, 1964). The results reported here are based on rankings of 10 personality traits contained in the questionnaire (which has been described in detail in the Porter & Henry article). These traits form two clusters and are intended to investigate a composite of the personality attributes underlying Riesman's (1950) Inner-Directed–Other-Directed distinction, Whyte's (1956) description of the Protestant ethic versus the social ethic, and the self-descriptions of top managers versus middle managers reported earlier by Porter and Ghiselli (1957). The 10 traits are listed below in the two theoretical clusters used as the basis for the analysis of the results:

Inner-Directed Cluster	*Other-Directed Cluster*
Forceful	Cooperative
Imaginative	Adaptable
Independent	Cautious
Self-Confident	Agreeable
Decisive	Tactful

Procedure and Sample

The questionnaire was distributed to approximately 800 commissioned officers and 700 noncommissioned officers throughout an Air Force overseas command. The sample was stratified to obtain proportionate representation from the hierarchical levels of grade and organization, and from the different functional activities comprising the command. Only officers occupying managerial-type positions and noncommissioned officers occupying supervisory-type positions were included. Usable responses were obtained from 1,297 individuals, or about 85% of the sample in both commissioned and noncommissioned ranks.

From personal-data questions asked on the last page of the questionnaire, respondents were classified by military grade since, among commissioned officers, grade correlates highly with level of managerial responsibility, and among noncommissioned officers it is the principal determinant of the supervisory hierarchy.

RESULTS

Comparisons between Military and Civilian Respondents

In this part of the Results section, comparisons will be made between the three levels of the commissioned officers sampled in the present study and three analogous levels of managers and executives from business and industrial firms sampled previously with the same questionnaire (Porter & Henry, 1964). As in the earlier comparative study of military and civilian managers based on these samples (Porter & Mitchell, 1967), the following pairings were considered as the most appropriate ones:

Brigadier generals and colonels	Vice-presidents
Lieutenant colonels and majors	Upper-middle managers
Captains and lieutenants	Lower-middle managers

The first six columns of Table 1 show the mean ranks for each of the 10 traits and the total of these for each cluster of traits for each of the three levels of military and civilian managers. The values in each cell in Table 1 for each trait have been subtracted from a constant number 10, so that larger numbers in Table 1 represent greater perceived importance.

As with the civilian managers, the cluster scores show that the military officers regarded the inner-directed cluster of traits as more important the higher the level of the organization, while the reverse was true of the other-directed cluster. (This reciprocal relationship must, of course, exist since the data are in terms of relative ranks rather than absolute ratings.) There is a far stronger increase associated with military rank, however, in the importance attached to inner-directed traits than was found among corresponding levels of civilian management. The total of the mean values assigned the inner-directed traits by senior as opposed to junior officers increases by 4.28, while the difference between corresponding levels of civilian managers is only 1.27.

Whereas all three groups of business managers clearly perceived the inner-directed group of traits as more important than the other-directed traits, the lowest ranking group of officers, the captains and lieutenants, assigned a slightly higher total to the other-directed cluster. Moreover, although the lieutenant colonels and majors valued inner-directed traits more highly than other-directed, their total for the inner-directed cluster was notably lower than that of their counterparts, the upper-middle business managers. The most senior military, the brigadier generals and colonels, however, ranked the inner-directed cluster somewhat higher than did their comparison group of vice-presidents, their cluster total being 26.50 as compared with 25.34 assigned by the vice-presidents.

In terms of the 10 individual traits, both samples show specific trait trends generally consistent with the overall cluster trends, although there are several interesting differences. The trend for Self-Confident in the inner-directed cluster is weak for civilian managers, and it is actually reversed for the military. That is, officers see Self-Confident as relatively *less* important to success with increasing rank. Tactful, in the other-directed cluster, is seen by officers as having increasing importance with higher rank, a perception directly opposite to that of the civilian managers.

One of the most interesting findings is the comparative ranking of Imaginative. Business and industrial managers consistently ranked this first in importance among the 10 personality traits. For the military, it assumed far less importance, being ranked sixth by both the junior and middle groups of officers and only fifth by the brigadier generals and colonels. Instead, Decisive assumed primary importance for the top- and middle-ranking military groups, while Self-Confident was ranked first by the captains and lieutenants. This latter group also attached great importance to Adaptable, ranking it second in overall importance. All other groups, military and civilian, perceived this trait as of far less importance.

Comparisons within the Military Hierarchy

Table 1 also presents results from the entire military sample and permits comparisons of the officers' rankings with those of the noncommissioned respondents.

The cluster totals for the inner-directed traits exhibit a continuous trend of increase in importance with higher rank throughout the military hierarchy. Thus, whereas commissioned officers, at least those in higher grades, strongly favor the inner-directed over other-directed traits, this preference is distinctly reversed for noncommissioned officers.

With respect to individual traits, all noncommissioned officers rank Self-Confident as the most important of the 10 personality traits for job success. Tactful is ranked second most important by the top- and middle-grade groups of noncommissioned officers and third by the staff sergeants. Interestingly, the importance of this trait increases within each of the two officer hierarchies, with the highest noncommissioned officer group placing more emphasis on it than any of the other five groups. Decisive, ranked as most important by the top and middle commissioned groups, drops to third for chief and senior master sergeants and to fifth for the other noncommissioned respondents.

TABLE 1

MEAN IMPORTANCE (10−$\bar{X}$ RANK) OF TRAITS FOR JOB SUCCESS: AIR FORCE OFFICERS VERSUS CIVILIAN MANAGERS

Trait	B.G./Col. (N = 22)	V. Pres. (N = 604)	Lt./Col./Maj. (N = 217)	Upper-Middle (N = 650)	Cpt./Lt. (N = 464)	Lower-Middle (N = 428)	Noncommissioned officers: CMS/SMS (N = 79)	MS/TS (N = 302)	SS (N = 213)
Inner-Directed									
Forceful	5.73	4.19	4.62	4.02	3.88	3.82	2.95	2.88	2.52
Imaginative	5.23	6.94	4.89	6.73	4.74	6.41	4.20	3.97	3.77
Independent	2.27	2.70	1.82	2.54	2.07	2.36	1.85	1.75	1.82
Self-Confident	5.86	5.43	5.84	5.36	5.86	5.87	6.55	6.73	6.73
Decisive	7.41	6.08	6.15	5.96	5.67	5.61	5.91	5.33	5.02
Total for cluster	26.50	25.34	23.32	24.61	22.22	24.07	21.46	20.66	19.86
Other-Directed									
Cooperative	5.09	5.49	5.69	5.61	5.70	5.88	5.90	6.09	6.41
Adaptable	4.32	4.77	5.37	5.13	5.81	5.16	4.88	5.37	5.48
Cautious	1.18	1.38	1.85	1.30	2.44	1.24	2.92	3.21	3.45
Agreeable	2.00	2.61	3.14	2.84	3.21	2.91	3.35	3.55	3.74
Tactful	6.04	5.42	5.64	5.52	5.61	5.74	6.47	6.11	6.04
Total for cluster	18.63	19.67	21.69	20.40	22.77	20.93	23.52	24.33	25.12

Note.—Higher numbers indicate greater importance. Abbreviations: B.G./Col. = Brigadier General/Colonel group; V. Pres. = Vice-President; Lt./Col./Maj. = Lieutenant/Colonel/Major group; Cpt./Lt. = Captain/Lieutenant group. CMS/SMS = Chief Master Sergeant/Senior Master Sergeant group; MS/TS = Master Sergeant/Technical Sergeant group; SS = Staff Sergeant group.

DISCUSSION

Perceptions of the behavior necessary for job success change from one level of the military management hierarchy to another, as is the case among business managers. However, the pattern of the change differs considerably from that exhibited by civilian executives. At the lower levels of the military, more organization-man-type behavior seems to be required than in equivalent levels in business organizations. Junior officers believe that to be successful they should demonstrate a somewhat more even blend of inner- and other-directed behavior than seems to be necessary in the business world. For senior military managers, however, the relative emphasis on inner-directed behavior clearly outweighs the importance of other-directedness.

A possible explanation of the junior officers' perceptions of successful behavior may be found in the nature of their managerial role in the Air Force. During these earlier years of their careers, they are both subject to more frequent reassignments and likely to experience widely differing types of managerial responsibility in successive assignments. Further, their duties are more likely to be circumscribed by procedural parameters than is the case for more senior officers—or for their junior-executive contemporaries in business and industrial management. In these circumstances, we do not find surprising the relatively stronger emphasis on traits such as Adaptable, Cautious, and Agreeable.

There are several other differences between the military and civilian managers that merit comment. While managers consistently ranked Imaginative as the most important among the 10 traits studied, officers placed this characteristic considerably lower and instead tended to identify Decisiveness and Self-Confident as the primary role requirements of those listed. Their emphasis on these two qualities is understandable. The officer as decision maker, "command-decisions," and similar concepts are an integral part of the role imputed to the military in our society. What is more important, these themes recur constantly in their continuing professional training and performance evaluation (Janowitz, 1960; Lang, 1965). The relatively low ranking of Imaginative is more difficult to explain. For some years the top leadership of the Air Force has repeatedly called for innovative and imaginative management by members of the officer corps.[1]

The results from the noncommissioned respondents show the clear dominance of other-directed behavior at these levels in the military hierarchy. Since the emphasis on this type of behavior decreases steadily at each higher level surveyed in the overall military hierarchy, as was true among business executives, it is interesting to speculate whether lower levels of supervision in industry and business would attach the same relative importance to other-directedness as do the Air Force noncommissioned officers. It must be pointed out, of course, that the noncommissioned supervisors in the sample obtained for this study were individuals whose occupational concern by definition is limited to administrative or logistical matters. Analysis of the perceptions of enlisted combat leaders, as in Infantry or Special Forces units, might yield quite different results.

[1] As, for example, in an address by (then) Secretary of the Air Force Eugene M. Zuckert to the Air War College, Maxwell Air Force Base, Alabama, on June 4, 1965.

REFERENCES

JANOWITZ, M. *The professional soldier.* Glencoe, Ill.: Free Press, 1960.

LANG, K. Military organizations. In J. G. March (Ed.), *Handbook of organizations.* Chicago: Rand McNally, 1965. Pp. 849–856.

PORTER, L. W., & GHISELLI, E. E. The self perceptions of top and middle management personnel. *Personnel Psychology,* 1957, 10, 397–406.

PORTER, L. W., & HENRY, M. M. Job attitudes in management: V. Perceptions of the importance of certain personality traits as a function of job level. *Journal of Applied Psychology,* 1964, 48, 31–36.

PORTER, L. W., & MITCHELL, V. F. Comparative study of need satisfactions in military and business hierarchies. *Journal of Applied Psychology,* 1967, 51, 139–144.

RIESMAN, D. *The lonely crowd.* New Haven: Yale University Press, 1950.

WHYTE, W. H., JR. *The organization man.* New York: Simon & Schuster, 1956.

(Received December 27, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 453–460

A PROJECTIVE STUDY OF ATTITUDES TOWARD CONTINUING EDUCATION [1]

IRWIN M. RUBIN AND HOMER G. MORGAN [2]

Alfred P. Sloan School of Management, Massachusetts Institute of Technology

A projective instrument aimed at measuring attitudes toward continuing education described a hypothetical research and development engineer as having (a) obtained an MS degree in a continuing education program, (b) completed 7 courses, or (c) completed only 1 course since receiving his BS. The 3 versions were randomly given to 312 engineers in a government research and development center. Research and development engineers did not perceive any significant difference between obtaining a degree or taking several courses in a continuing education program. However, obtaining a degree or taking several courses were both seen as having more positive attributes than taking only 1 course, were perceived as being associated with more management potential, higher ambition, and greater professionalism, and as keeping an engineer more up to date.

Continuing education and technical obsolescence of engineers have been subjects of much discussion. Numerous reports have been written and many conferences have been held. All of the studies on technical obsolescence conclude that technical learning is necessary throughout the lifetime of an engineer.[3] They suggest that the current pace of technology requires that informal, on-the-job type of learning must be supplemented by more formal continuing educational opportunities. These opportunities can take many forms, as pointed out in a recent study by the Staff of the Industrial Research Institute (1965), ranging from attendance at technical meetings and seminars, through participation in company-sponsored classes, to leave for study as resident graduate students. However, the content of most of the available literature was indicated by Foecke (1966) in a recent article when he said: "One need be neither very perceptive nor snobbish to say that the body of literature on continuing education for engineers is not characterized by a wealth of penetrating discussion [p. 881]."

With few exceptions, little empirical research, as opposed to "armchair philosophizing," has been conducted in this area. Cenko (1964) collected questionnaire data from engineers and engineering supervisors in three large research and development corporations which indicated that, as expected, obsolescence increased the longer a person was out of school.[4] He also found that engineers who completed at least two advanced engineering courses in a 4-yr. period were perceived as having obsolesced less than engineers who did not take courses. Another result of his investigation was that engineers with above-average mathematical ability, first, are more likely to pursue and complete advanced engineering courses, and, second, have a better opportunity for improving or maintaining their technical skills than other engineers. This work has substantiated the assumption of most authors that continuing education (in this case, advanced engineering courses) is a valid technique for combating technical obsolescence. Engineers themselves regard technical courses as important or very important to their professional development, as over 80% reported in one survey (Anonymous,

[1] The data reported in this paper were collected in conjunction with an unpublished master's thesis by Morgan (1967). The work was done in part at the Massachusetts Institute of Technology (M.I.T.) Computation Center.

[2] Presently with the Langley Research Center of the National Aeronautics and Space Administration (NASA).

[3] See, for example, American Society for Engineering Education (1964); Engineers Council for Professional Development (1966); National Society of Professional Engineers (1965); Torda (1963).

[4] This study defined an "obsolescence index" as the accumulated scores on a series of questions answered on a 7-point scale. A weighted average of an engineer's responses and his supervisor's ratings gave the final index.

1964). Still another study by Warkov and Marsh (1965) revealed that recipients of any kind of supplementary training were in a minority in 1962. Only 41% of the engineers they surveyed had participated in any kind of continuing education. They also found that supplementary training was an inverse function of formal educational attainment—the higher the degree, the less likelihood of having participated in supplementary training.

These few empirical studies discussed above have suggested that continuing education is highly desirable for engineers and that engineers themselves recognize the desirability. Still, only a minority of such personnel have actually participated in any supplementary training. However, the problem of motivating the individual to continue his education has previously been recognized. For example, a recent conference (Torda, 1963) pointed out:

> One thing is clear: continued education of a graduate is imperative and it should be supported by both industry and colleges. Motivating an individual to continue his education is another difficult problem and economic rewards as well as recognition of the importance of engineers and their performance seem to be equally vital in achieving this goal [p. 3].

In fact, attendees at this same conference felt that the attitudes of individuals were a major factor in obsolescence, as they reported, "Technical obsolescence is always exclusively a personal problem of the engineer's attitude [p. 98]."

Since it has been suggested that continuing education is beneficial in alleviating technical obsolescence, but that participation is mainly a function of the individual's attitude, this study attempted to learn more about attitudes of engineers toward continuing education programs. It was restricted to research and development personnel who would appear to require constant updating to stay abreast of their rapidly changing technology. Specifically, this study was designed to test two general hypotheses: first, research and development engineers regard continuing education as a means of avoiding technical obsolescence, and, second, research and development engineers regard continuing education as a method for organizational advancement.

METHOD

Measurement Instrument

The nature of the phenomenon under investigation—attitudes toward continuing education—and the sophisticated level of the relevant subject population—engineers and scientists—led us to the development of a projective-test instrument. The projective method has been described by Mason Haire (1950) in his classic paper on application of projective techniques to market research as:

> Basically, a projective test involves presenting the subject with an ambiguous stimulus—one that does not quite make sense in itself—and asking him to make sense of it. The theory is that in order to make sense he will have to add to it—to fill out the picture—and in so doing he projects part of himself into it. Since we know what was in the original stimulus we can quite easily identify the parts that were added, and, in this way, painlessly obtain information about the person [p. 650].

Projective methods, in other words, are useful whenever people are unwilling or unable to discuss their true attitudes. The projective-test method makes it easier for the person to express himself since he is not talking explicitly about his own attitudes and feelings. He may also be unable to describe his own feelings and attitudes as accurately as they can be discerned in the projective-test situation. This method was thought to be particularly applicable to the present study since it was desired to obtain the attitudes of engineers toward a subject (continuing education) which intimately affected them and their associates.

The final projective instrument took the form of a resumé of a research and development engineer,[5] as follows:

> The personnel department is considering hiring an engineer to assist you in your projects. The following resumé has been received from a potential new employee and you are to evaluate his suitability as your assistant. Please read it carefully.

RESUMÉ

Vital Statistics:
  Name: John Doe
  Date of Birth: June 4, 1936
  Height: 5′11″ Weight: 175 lbs.
Marital Status:
  Married, two children
Military Service:
  None

[5] An earlier version of this instrument took the form of single-paragraph descriptions of Joe Doe, an aerospace engineer working in a NASA Research Center. The results of a pretest brought to light many problems. The idea of using a resumé was suggested by a colleague at M.I.T., Tom Allen. The value of a pretest will never be more clear to us since it is doubtful if usable data would have been obtained with the original version of the test instrument.

Education:

B.S. in Aerospace Engineering, University of Michigan, 1958

M.S.[6] in Aerospace Engineering, Case Institute of Technology, 1962 (obtained through company sponsored graduate training program)

Health:

Good

Professional Experience:

NASA Lewis Research Center, Cleveland, Ohio, 1958 to present

Research and Development projects

Experimental equipment design

Data analysis and processing

Supervised laboratory technicians

In order to allow the subject to project himself into the resumé, the fictitious engineer was presented as having worked at the NASA Lewis Research Center, which is a sister installation of the NASA center where the data were collected. The educational background was made both neutral and realistic by selecting a good but not top engineering school for his alma mater and using a good technical school near the Lewis Research Center for his continuing education (Walsh, 1966).

The projective stimulus was presented to an individual without an explanation of its true intent. The instructions said it was a study of the perceptions of research and development engineers and asked the participant to assume that he had been asked by his personnel department to evaluate the resumé of a potential new employee. In order to avoid any implied rivalry, it was suggested that the new employee was to be hired as the participant's assistant. No mention of continuing education or graduate extension courses was made except for the cues listed under the educational qualifications of the resumé.

The respondent was asked to make his evaluation by rating the engineer described in the resumé on a series of bipolar adjective pairs. The majority of these adjective pairs were adopted from prior research by Schein (1965) in his studies of student's images of their professors. The items were meant to tap such factors as activity, competence, maturity, and individualism. Several pairs, such as "obsolete–up to date," were added because of their particular relevance to the present study. Several control items, such as "warm–cold" whose scoring should not be influenced by attitudes toward continuing education, were also added.

From a total of 43 adjective pairs, 12 items were selected[7] for analysis whose perception was thought to be most dependent upon the version of the projective stimulus seen. The selected items were ambitious–unambitious, high initiative–low initiative, intelligent-nonintelligent, management potential–not management potential, professional–unprofessional, scientific–unscientific, active–passive, growing–stationary, up to date–obsolete, clear thinking–muddled thinking, enthusiastic–unenthusiastic, and aggressive–timid. In addition, as a control, three adjective pairs (tolerant–prejudiced, warm–cold, and poised–awkward) were included for analysis. The analysis of these three items, which should *not* be perceived differently due to exposure to the three stimuli, was expected to add confidence to any results obtained. For discussion purposes, the polarity of some of the adjective pairs has been reversed from their polarity on the test instrument in order to make the more "desirable" or "positive" state appear on the left pole.

These adjective pairs would enable us to test the general hypothesis that the engineers described by the "MS" and "seven-course" stimuli would be seen in a more positive light than their counterparts in the "one-course" condition.

The Study Group

The study group was selected from the research and development engineers at the Langley Research Center of the National Aeronautics and Space Administration. Langley has about 4,300 employees, of whom approximately 1,500 are engineering and scientific professionals. Most of the professionals are assigned to 10 major research divisions which are organizationally separate from groups responsible for administration and technical support functions. The personnel surveyed in the present study consisted of all professionals classified as aerospace technologists, excluding supervisors, in 4 of the 10 research divisions.[8]

An interview with a member of the Training Office staff at Langley indicated the pervasiveness of the continuing education program. Each year, about one-

[6] This is the "MS" version. The "seven-course" version read: "7 graduate courses in aerospace engineering, Case Institute of Technology, 1959–1965." The "one-course" version read: "1 graduate extension course in aerospace engineering, Case Institute of Technology, 1959."

[7] The initial plan was to factor analyze the 43 adjective pairs. This process failed to reveal any clear-cut factors. The experimental items were selected *prior* to any further data analysis.

[8] The participating groups have an extensive continuing education program available to them. Since Langley is a government center, its educational programs are based on Public Law 85-507, the Government Employees Training Act of 1958, which established a policy of encouragement and support for continuing education of government employees. Langley implements the policy by offering extension courses in technical subjects, sponsored by one of three nearby universities, during duty hours with all fees for qualified employees paid by the government. Professional employees who have been accepted for graduate study and have successfully completed two extension courses are also eligible for graduate-study leave for full-time resident study, whenever such study will upgrade their capabilities. While on graduate-study leave, an employee's salary and tuition fees are paid by the government. Resident study such as this usually leads to a graduate degree with the thesis research carried out in conjunction with a Langley project.

third of the professional staff is active in the program. It is estimated that 90% of the research professionals have participated at one time or another, and this year (1967) employees are studying at 27 different institutions. In the last 15 yr., 257 master's degrees and 28 doctoral degrees were earned from 33 institutions. In terms of demographic characteristics, the average age of the group was 33.9 yr., with extreme values of 22 and 63 yr. The average experience level was 9.8 yr. while the average length of agency service was 9.2 yr. With respect to education, 2% had doctorates, 25% had master's degrees, and 73% had bachelor's degrees. Fourteen percent considered themselves to be in analytical work, 31% were engaged in experimental work, and 55% performed a combination of analytical and experimental work.

Participation in their agency's continuing education program was quite varied among the group surveyed. Thirty-three percent had been on graduate-study leave at some time in their careers—that is, they had been away from their jobs and had been full-time students. Twelve percent of the group had received a graduate degree in the continuing-education program while 32% were currently candidates for graduate degrees. Of those who were not degree candidates and had not earned a degree in the program, 13% had taken five or more courses, 21% had taken three or four courses, and 30% had taken one or two courses. Only 22% of the complete group had not participated in the program in any way.

Data-Collection Procedures

The four research divisions surveyed were selected as being representative of the complete center and large enough to provide an adequate number of responses. Lists of personnel assigned to the four divisions were supplied by the Personnel Division. This list was edited by removing all supervisory personnel (those having line responsibility in the official organizational structure), all nonprofessional employees, and all professionals on graduate-study leave. The three versions of the test instrument were distributed *randomly*[9] by alternating the three versions with the personnel list arranged in alphabetical order.

The test instrument was sent to 370 individuals through the intercenter mail. Only responses received within 10 days were included in the subsequent study, but 312 usable replies, or 84% of the questionnaires mailed, were received before the cutoff. Of the 312 responses, 108 had the MS-degree version of the test instrument, 104 had the seven-course version, and 100 had the one-course version. Only 5 responses were received which were unusable. Anonymity was maintained throughout the study.[10]

RESULTS

The original hypotheses stated that the engineer described in the MS condition would be seen as having more positive traits than either his seven-course or one-course counterparts. In addition, the engineer described in the seven-course condition would be seen more positively than his one-course counterpart. Finally, these three hypothetical engineers would be seen as no different on the traits of tolerant–prejudiced, warm–cold, and poised–awkward.

The results are presented in Tables 1, 2, and 3 and are summarized graphically in Figure 1. The tables contain three sets of information about the groups which they compare. First, the median scores for the ratings of each group are presented (the lower the score, the more positive the rating). Next, the tables contain the percentage of each group which scored above the combined group median. Finally, the significance of the difference between the medians, as determined from a Mann-Whitney U test, is provided.

The first conclusion which can be drawn from these data is that, within the sample studied, no differences exist between the MS and seven-course conditions. (See Table 1.) A person receiving a master's degree is seen as no more ambitious, intelligent, up to date, etc., than an engineer who has completed seven advanced courses. The first of our hypotheses is therefore unsubstantiated.

With respect to the other two hypotheses, the results are remarkably consistent and clear. On each of the 12 experimental items, with a very few exceptions, the MS group and seven-course group are seen more positively than their one-course counterparts. (See

[9] The success of this process was checked and it was found that each of the three experimental groups contained people who were virtually identical in terms of age, professional experience, years with NASA, type of work performed, number of previous courses taken, and degree of participation in the Graduate Training Program.

[10] Since the test instrument is a projective device, its success in disguising the intent of the study is of interest. Several notes were included with the returned questionnaires which indicated that the test instrument was sufficiently ambiguous. Perhaps the most informative feedback came from one very serious participant who telephoned to complain about the request for an evaluation based on such flimsy data. He ended with the advice that the resumé was so vague that the questionnaires would actually reveal more about the people filling them out than about the engineer being evaluated!

TABLE 1

COMPARISON OF MEDIAN RESPONSES TO ADJECTIVE PAIRS FOR MS-DEGREE VERSUS SEVEN-COURSE CONDITIONS

Adjective pair	*Mdn*		Percentage above group median		p[a]
	MS-degree group (N = 106)	Seven-course group (N = 99)	MS-degree group	Seven-course group	
Ambitious–unambitious[b]	3.00	3.30	53.8	46.1	.23379
High initiative–low initiative	3.52	3.69	53.2	46.6	.14279
Intelligent–nonintelligent	3.39	3.67	54.9	44.9	.06901
Management potential–not management potential	3.48	3.82	55.5	44.3	.07374
Professional-unprofessional	3.07	3.33	54.1	45.7	.09475
Scientific–unscientific	2.81	2.92	52.4	47.6	.15528
Active–passive	3.41	3.35	49.2	50.8	.34890
Growing–stationary	3.46	3.49	50.4	49.5	.33720
Up to date–obsolete	3.57	3.50	48.8	51.2	.39946
Clear thinking–muddled thinking	3.74	3.70	49.5	50.5	.33641
Enthusiastic–unenthusiastic	3.84	3.74	48.7	51.3	.35125
Aggressive–timid	3.88	4.03	52.6	47.4	.35195
Tolerant–prejudiced	4.14	4.20	51.5	48.4	.33558
Warm–cold	4.37	4.36	49.7	50.3	.43642
Poised–awkward	4.69	4.75	48.4	51.7	.27520

[a] Mann-Whitney U test on difference of medians, one-tailed.
[b] The MS-degree group is seen as more ambitious *relative* to the seven-course group.

TABLE 2

COMPARISON OF MEDIAN RESPONSES TO ADJECTIVE PAIRS FOR MS-DEGREE VERSUS ONE-COURSE CONDITIONS

Adjective pair	*Mdn*		Percentage above group median		p[a]
	MS-degree group (N = 106)	One-course group (N = 94)	MS-degree group	One-course group	
Ambitious–unambitious[b]	3.00	4.08	61.2	37.9	.00021
High initiative–low initiative	3.52	4.34	63.2	35.8	.00006
Intelligent–nonintelligent	3.39	4.18	62.0	37.1	.00022
Management potential-not management potential	3.48	4.08	60.0	39.1	.00097
Professional–unprofessional	3.07	3.54	57.9	41.6	.00820
Scientific–unscientific	2.81	3.29	58.6	40.8	.01530
Active–passive	3.41	3.93	56.6	42.8	.01843
Growing-stationary	3.46	3.97	58.4	40.9	.00086
Up to date–obsolete	3.57	4.25	62.1	37.0	.00017
Clear thinking–muddled thinking	3.74	4.12	55.6	44.0	.12578
Enthusiastic–unenthusiastic	3.84	4.25	55.6	43.9	.00955
Aggressive-timid	3.88	4.11	53.8	45.8	.20851
Tolerant–prejudiced	4.14	4.20	51.6	48.3	.30300
Warm–cold	4.37	4.39	50.4	49.5	.44490
Poised–awkward	4.69	4.66	51.5	48.4	.32712

[a] Mann-Whitney U test on difference of medians, one-tailed.
[b] The MS-degree group is seen as more ambitious *relative* to the one-course group.

TABLE 3

COMPARISON OF MEDIAN RESPONSES TO ADJECTIVE PAIRS FOR SEVEN-COURSE VERSUS ONE-COURSE CONDITIONS

Adjective pair	*Mdn*		Percentage above group median		p[a]
	Seven-course group ($N = 99$)	One-course group ($N = 94$)	Seven-course group	One-course group	
Ambitious–unambitious[b]	3.30	4.08	61.2	38.5	.00042
High initiative–low initiative	3.69	4.34	60.9	38.6	.00177
Intelligent–nonintelligent	3.67	4.18	58.4	41.4	.00834
Management potential–not management potential	3.82	4.08	53.6	46.2	.05466
Professional–unprofessional	3.33	3.54	53.8	46.1	.13858
Scientific–unscientific	2.92	3.29	55.9	43.9	.13241
Active–passive	3.35	3.93	57.9	41.8	.00519
Growing–stationary	3.49	3.97	58.3	41.6	.00376
Up to date–obsolete	3.50	4.25	63.5	36.1	.00011
Clear thinking–muddled thinking	3.70	4.12	56.2	43.5	.04306
Enthusiastic–unenthusiastic	3.74	4.25	57.0	42.7	.00250
Aggressive–timid	4.03	4.11	51.2	48.7	.33703
Tolerant–prejudiced	4.20	4.20	50.1	49.9	.44164
Warm–cold	4.36	4.39	50.7	49.2	.49625
Poised–awkward	4.75	4.66	53.0	46.9	.18286

[a] Mann-Whitney U test on difference of medians, one-tailed.
[b] The seven-course group is seen as more ambitious *relative* to the one-course group.

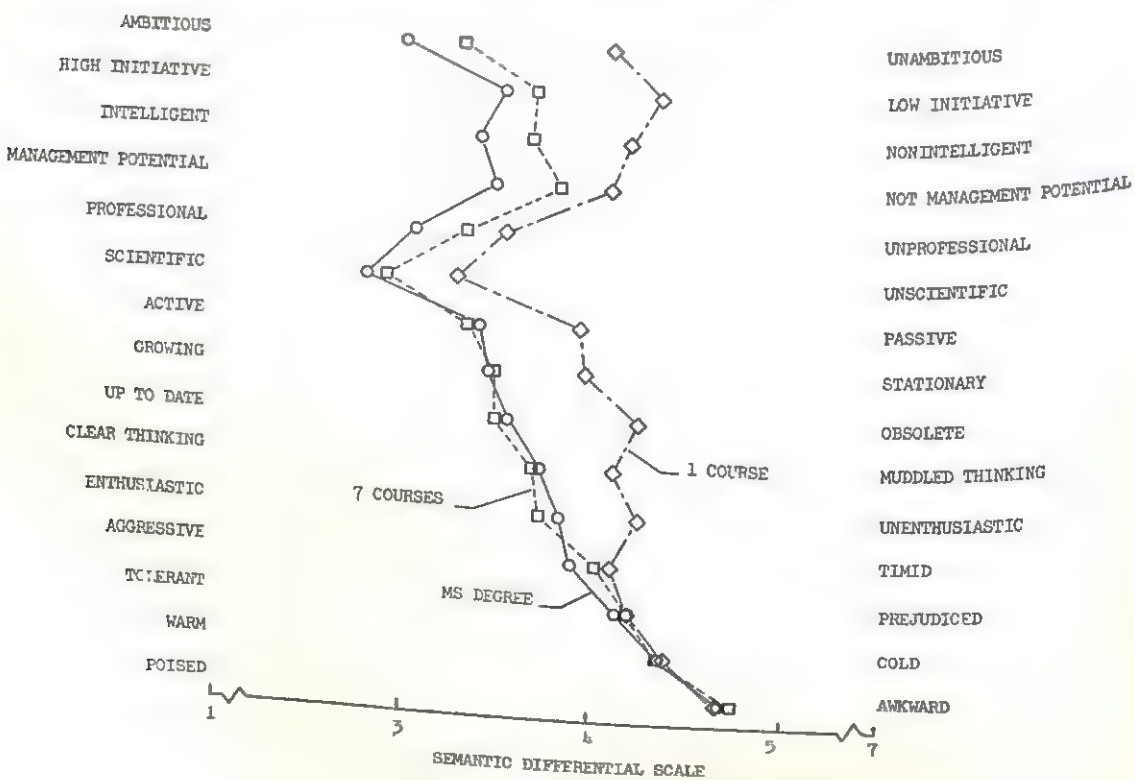


FIG. 1. Median score profiles.

Tables 2 and 3.) Furthermore, as hypothesized, no differences are observed on the three adjective pairs which were included as controls. The exception is most consistent with respect to aggressive–timid. In the other cases where the differences do not reach statistical significance (Table 2, clear thinking–muddled thinking, and Table 3, scientific and active), the difference is in the predicted direction.

These results are all the more striking when one considers one other factor. It will be recalled that data were collected on 43 adjective pairs. The 15 items used for hypothesis testing were selected *prior* to any data analyses. After testing the hypotheses, the responses to 24 of the remaining adjective pairs across the three experimental conditions were compared. Of these 72 pair-wise comparisons (24 items across 3 groups compared 2 at a time), 4 reach statistical significance. This is almost the exact number of significant differences (3.6 or 72×0.05) one would expect to find by chance alone!

Discussion

Within the sample of engineers studied, it is clear that certain very strong opinions are held with respect to continuing education. A closer look at the adjective pairs gives an idea about specific attitudes. If the MS-degree and seven-course groups are lumped together, for the moment, as the continuing-education group, Figure 1 shows that participation in continuing education is perceived as making an engineer less obsolete. He is regarded as clearer-thinking, growing, active, and more up-to-date. Continuing education is also perceived as a mark of professionalism—a participating engineer is seen as more intelligent, professional, and scientific than a nonparticipating engineer. A participating engineer is also regarded as having more ambition, higher initiative, more enthusiasm, and higher management potential than a nonparticipating engineer.

Contrary to initial expectations no significant differences were found between the MS and seven-course conditions. It may be that the engineer described in the seven-course condition was seen as working toward but not having yet received an advanced degree. On the other hand, the actual degree may be less important than the steps necessary to get the degree, that is, taking many courses. One way to examine this issue is to widen the gap between two such hypothetical engineers. In other words, how will peoples' perceptions of an engineer who has earned an MS degree 8 yr. ago and taken no courses since then differ from an engineer who has taken one course per year since receiving his BS but has not received an advanced degree.

It is recognized that many factors can be expected to influence the attitudes of research and development engineers toward continuing education. Within the design utilized in this study, the variable effects of such factors as age, experience, and educational level were eliminated through randomization. Further research could fruitfully focus on the relationship between individual characteristics and attitudes toward continuing education.

The fact that our results strongly suggest that engineers perceive continuing education as contributing to managerial advancement raises interesting questions about the reward structure in an organization. The issue revolves around the dual-ladder concept of movement through an organization's hierarchy. To what extent does an organization want its engineers to view continued technical education as a path to promotion into the managerial hierarchy? It is conceivable that some engineers might not bother with continuing education because the reward of becoming a manager is not what they desire. Many have a desire to continue in their technical specialties, and it is to the organization's benefit to make certain these people are rewarded in a manner consistent with their interests.

The organization we have studied is one in which continuing education is heavily endorsed, as evidenced by the opportunities made available to people and the number of people who take advantage of these opportunities. It would be very interesting to study the relationship between the climate in an organization with respect to continuing education and individual attitudes. In what ways, if any, would peoples' attitudes have been different if we had studied an organiza-

tion which provided very few opportunities for continuing education.

Finally, we are very much encouraged by these initial results achieved through the utilization of a short, inexpensive, easy-to-administer, projective instrument. The effects of many different stimuli can be investigated within the framework of the method we used. With very few changes, one could study attitudes toward short courses versus graduate-level courses, in-house versus on-campus courses, management subjects versus technical subjects, and a variety of other important issues related to the question of technical obsolescence and the role of continuing education.

REFERENCES

AMERICAN SOCIETY FOR ENGINEERING EDUCATION. Goals of engineering education. (Information Document No. 4) Initial report to the Institutional Committees Concerning Graduate Education in Engineering. Wash., D. C.: ASEE, 1964.

ANONYMOUS. Who pays for technical retooling? *Machine Design*, July 2, 1964, 92–95.

CENKO, S. Factors in obsolescence of engineering knowledge. Unpublished master's thesis, Sloan School of Management, Massachusetts Institute of Technology, 1964.

ENGINEERS COUNCIL FOR PROFESSIONAL DEVELOPMENT. Continuing engineering studies—a report of the Joint Advisory Committee. Washington, D. C.: ECPD, 1966.

FOECKE, H. A. Continuing education for engineers. *Science*, 1966, **152**, 880–883.

HAIRE, M. Projective techniques in market research. *Journal of Marketing Research*, 1950, **14**, 649–656.

MORGAN, H. Attitudes of research and development engineering toward continuing education. Unpublished master's thesis, Sloan School of Management, Massachusetts Institute of Technology, 1967.

NATIONAL SOCIETY OF PROFESSIONAL ENGINEERS. Engineers and management look to the future. In, *Proceedings of the Second Annual Conference of Professional Engineers in Industry*. (No. 1429) Detroit: NSPE, 1965.

SCHEIN, E. H. The student image of the teacher in a business school. Sloan School of Management Working Paper No. 123-65, Massachusetts Institute of Technology, May 1965.

STAFF OF THE INDUSTRIAL RESEARCH INSTITUTE. Continued development of technical personnel—prevention of obsolescence. *Research Management*, 1965, **8**(3), 159–167.

TORDA, T. P. (Ed.) *Proceedings of the Midwest Conference on Reducing Obsolescence of Engineering Skills*. Chicago: Illinois Institute of Technology, March 26–27, 1963.

WALSH, J. Graduate education: ACE study rates departments qualitatively. *Science*, 1966, **152**, 1226–1228.

WARKOV, S., & MARSH, J. The education and training of America's scientists and engineers. (Report No. 104) Chicago: National Opinion Research Center, University of Chicago, October 1965.

(Early publication received June 12, 1967)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 461–468

SELECTION INTERVIEW DECISIONS:

THE RELATIVE INFLUENCE OF APPEARANCE AND FACTUAL WRITTEN INFORMATION ON AN INTERVIEWER'S FINAL RATING

ROBERT E. CARLSON

Life Insurance Agency Management Association, Hartford, Connecticut

Previous research by the author has shown that life insurance agency managers serving as interviewers report they will make employment decisions on the basis of appearance. In addition, it has also been shown that these same managers report they will make employment decisions on the basis of factual written summaries of an applicant. However, previous research has not investigated the relative effect of appearance vs. factual written data on an interviewer's final evaluation of an applicant for the job of life insurance agent. It was the purpose of this research to study these relative effects. The appearance and written information were presented to Ss in complementary and contrasting patterns. It was found that the appearance data had little impact on the final rating. It was also found that when the appearance data and written information were presented in a complementary manner there existed a component in the final rating due to information favorableness greater than that contributed by the separate ratings of the appearance and written information.

Recent studies of the selection interview have concentrated on the decision-making process and the variables which influence it. From such a study, Springbett (1953) concluded that the interview is primarily a search for negative information. A later study by Bolster and Springbett (1961) confirmed this finding and, in addition, concluded that the interviewers most ready to commit themselves to a decision were the quickest to return to a noncommittal position when information of opposite value was presented. More recent studies have presented additional evidence that negative information is given more weight in the final decision than is positive information (Mayfield, 1964; Mayfield & Carlson, 1966).

These and other studies (Webster, 1964) have indicated that future research on the decision-making process involved in the selection interview will become more complex. For example, the studies noted earlier varied the favorableness of the information presented to various interviewers. Yet Mayfield and Carlson (1966) have since shown that properties other than the favorableness of the specific items of information about an applicant might be related to the outcome and the reliability of the final decision. In addition to favorability, these properties were: (1) the interrater agreement on the items of information and (2) the intrarater agreement of the items.

Another information property which has been shown to have a differential effect on the outcome of simulated interviews is the objectivity of the information presented to interviewers (Carlson, 1967a). For example, in previous studies the author has found that in the absence of other information, life insurance agency managers acting as interviewers reported they would offer an average of 1.83 employment contracts for a sample of eight applicants on the basis of appearance alone (Carlson, 1967a; Carlson & Mayfield, 1967). In the same studies using a second sample of managers as interviewers, they reported they would make an average of 2.60 selection decisions on the basis of written information alone. The mean difference between the number of applicants reportedly hired from photographs and the number hired from written information was statistically significant at the .01 probability level. The implication was that managers were more influenced by the objective written information than they were by the more subjective appearance data. However, the design

of the previous study did not allow a relative comparison between the reported use of appearance data and the use of written information. It is the purpose of this study to allow interviewers to use both appearance data and written information and to examine the relative effect of each on the final evaluation of the applicant, while controlling on the favorability, inter-, and intrarater agreement of both types of input information. No a priori hypotheses were advanced as to the relative strength of appearance versus written information or the effect of the various information properties on the final interviewer ratings.

METHOD

Subjects

The responding Ss who played the role of interviewers were life insurance agency managers who were attending or who had attended 2-wk. schools in agency management conducted by the Life Insurance Agency Management Association. As managers, they have complete responsibility for recruiting, interviewing, selecting, and training salesmen to work under their supervision.

Construction of Hypothetical Applicants

Using a paper-and-pencil approach, lists of 125 individual items of information selected at random from a total pool of 600 items were administered to life insurance agency managers attending 2-wk. schools. The reasons for the paper-and-pencil approach and the method of constructing the individual items of information have been described in detail elsewhere (Mayfield & Carlson, 1966).

TABLE 1

IDENTIFICATION OF HYPOTHETICAL APPLICANTS IN TERMS OF PSYCHOMETRIC PROPERTIES OF INFORMATION USED TO CONSTRUCT THEM

Hypothetical applicant	Measure		
	Favorableness	Interrater agreement	Interrater agreement (stability)
1	High	High	Low
2	Low	High	High
3	High	High	High
4	High	Low	Low
5	Low	Low	Low
6	Low	High	Low
7	Low	Low	High
8	High	Low	High

Note.—High: above the median in the respective distribution. Low: below the median in the respective distribution.

Each S was instructed to consider each item in isolation and rate it on a 7-point scale as to how favorable it would be if it were true about an applicant he was considering hiring to work under his supervision. The S also had an opportunity to check one of two boxes which indicated that this item by itself would either completely qualify the applicant for employment or disqualify him from further consideration.

Each of the 600 items was evaluated by over 100 managers, and the item means and variances computed. The mean served as a measure of item favorableness, and the variance served as a measure of interrater agreement. One month after the manager had rated the items at the school, the same questionnaire with the items rearranged was mailed to each S. On the basis of the return (over 75%), the percentage of perfect agreement measure (Brown, Lucero, & Foss, 1962) for each item was calculated and used as a measure of the item's intrarater (test-retest) agreement. This measure is simply the percentage of respondents who rated the item exactly the same on both administrations, and is particularly useful when dealing with items having restricted or skewed distributions.

The three univariate distributions of 600 means, variances, and test-retest measures were each dichotomized at the median. Each item was then classified as being either above or below the median in favorableness, interrater agreement (variance), and intrarater (test-retest) agreement. Thus, each of the 600 items fell into one of eight cells representing a 2^3 factorial arrangement. From the items in each cell a hypothetical applicant was constructed consisting of 30 items. The only departure from random selection of items occurred when two of the same type of items were chosen, such as two different ages for the same applicant. Thus, each of the eight hypothetical applicants was made up of nonoverlapping, consistent, and homogeneous information, that is, information with similar psychometric characteristics. For example, one applicant was made up of items which had high favorability, low interrater agreement, and high intrarater agreement. A second applicant consisted of items having exactly the opposite values, while a third applicant was constructed of items having high favorability, low interrater agreement, and low intrarater agreement, a fourth, etc. Table 1 presents the makeup of all eight hypothetical applicants.

A procedure similar to that used in constructing hypothetical candidates from personal-history items was employed to develop psychometric values for photographs. The original pool of photographs consisted of 39 photographs of Life Insurance Agency Management Agency personnel. A development sample of 315 managers (a different sample of managers from those who rated the written items) rated each photograph on a seven-place predicted behavior rating scale. From the ratings and a follow-up study, the photograph mean, variance, and test-retest reliability were computed. The ratings of the photographs were also submitted to a factor anal-

ysis in an attempt to eliminate those photographs which were evaluated on the basis of quality or some general response set.

Based upon the item statistics and the results of the factor analysis, eight photographs were selected, one each for the eight cells formed by dichotomizing the distribution of photograph means, variances, and measure of intrarater (test-retest) agreement. The selection of the photographs represented the same 2^3 factorial arrangement as did the written information.

The photographs and the hypothetical applicants represented by written information were combined in complementary and contrasting patterns. In the complementary pattern, the photograph and written information were of the same relative level. That is, a photograph with high favorability, low interrater agreement, and low intrarater agreement was matched with the applicant represented by 30 items of written information all of which received highly favorable ratings with low inter- and intrarater agreement. Thus, for Sample 1, the photograph and written information were of a complementary value.

The contrasting or cross-over pattern grouped a photograph with written information of opposite value. A photograph with high favorability and low inter- and intrarater agreement was paired with an applicant represented by written information of low favorability and high inter- and intrarater agreement. Thus, for Sample 2, the photograph and written information were of contrasting values.

As a result of the above two methods of matching, two, 2^3 factorials were constructed. A test booklet was prepared representing each of the 16 cells. The first page of the booklet contained general information and directions. The first page was followed by a blank page. The third page contained the photograph; the fourth page was blank; the fifth page the written information; the sixth page was blank; the last page contained the photograph and written information combined. For each of the three correlate situations (photograph, written information, photograph and written combined) the managers were to rate the prospective applicant on a nine-step behavior description rating scale indicating relative potential for success as a life insurance agent, make an employment decision to select, reject, or remain undecided for the applicant; indicate on a five-position scale how the hypothetical applicant measured up against applicants the manager had interviewed the past year. The managers were requested not to look ahead and there is no evidence that they did. Each manager rated only one complete hypothetical applicant. One month after the initial administration, the same booklet was mailed to each *S* for test-retest purposes. The stability of the ratings are the subject of another study and are reported elsewhere (Carlson, 1967b). The sample sizes are presented in Tables 2 and 3.

Analysis

The design of the study consisted of two, 2^3 factorials, each with two covariates. The treatment combinations were assigned to *S*s randomly. The departure from equal cell frequencies is due to incomplete information or nonreturns. The three independent variables were two levels of information favorableness and two degrees each of inter- and intrarater agreement. The dependent variable was the rating of the photograph and written information combined. The covariates were the separate ratings of the photograph and written information.

The analysis consisted of a "step-wise" analysis of covariance using the general linear hypothesis model. The first step in the analysis was a 2^3 analysis of variance which provided comparative data. The second step consisted of treating the rating of the photograph alone as a covariate in an analysis of covariance within the 2^3 model. The third step treated the rating of the written information alone as a covariate within the 2^3 model. The fourth step used the separate ratings of the photograph and the written information both as covariates. The dependent variable for each step in the analysis was the rating of the photograph and written information combined. By comparing the adjusted mean squares at the end of each step, the effect of the covariate may be examined. The adjusted mean square at the end of the fourth step represents some component of the interviewer's rating greater than that contributed by the best linear combination of the separate rating of the photograph and the written information, that is, some error or "clinical component" the interviewer can manufacture from the combined presence of the photograph and written data which is different from the separate ratings of the photograph and written information.

The analytic procedure was repeated for both the complementary and contrasting pattern of hypothetical applicants. The original analysis had planned to treat the study as a 2^4 analysis of covariance, with complementary versus contrasting patterns as the fourth independent variable. However, as can be seen from Tables 2 and 3, the regressions are more homogeneous within pattern of presentation than across the pattern of presentations. In addition, Tables 2 and 3 indicate that although the multiple correlations between the combined rating and the separate ratings of the photograph and written information are quite high, they seldom account for greater than 50% of the variance in the overall rating. Thus, there appears to be substantial variance above and beyond that accounted for by the separate ratings and which may be differentially effected by the independent variables. The results of the analysis of variance and covariance are presented in Tables 4 and 5 while Tables 6 and 7 present the observed and adjusted means.

RESULTS

Complementary Information

Table 4 presents the results of the four-step analysis for applicants constructed of complementary information. Looking first at

TABLE 2

STANDARD NORMAL REGRESSION WEIGHTS AND MULTIPLE CORRELATIONS FOR APPLICANTS CONSTRUCTED OF COMPLEMENTARY INFORMATION

Degree of favorability	Low intrarater stability		High intrarater stability	
	Low interrater agreement	High interrater agreement	Low interrater agreement	High interrater agreement
Unfavorable				
B_2	.23	.02	.11	−.13
B_3	.73	.78	.57	.67
$R_{1.23}$	.73	.78	.61	.67
N	38	58	56	43
Favorable				
B_2	.36	.38	−.08	.39
B_3	.70	.67	.81	.59
$R_{1.23}$	.74	.80	.83	.79
N	41	47	55	45

Note.—Abbreviations: B_2 = standard normal regression coefficient, rating of photograph; B_3 = standard normal regression coefficient, rating of written information; $R_{1.23}$ = multiple correlation. Rating of photograph and written information combined is dependent variable, rating of photograph alone and written information alone as independent variables; N = sample size.

the effects of the covariates on the error mean square, it can generally be concluded that the effect of the photograph on the final judgment is negligible, while the effect of the written information is substantial. Specifically, treating the rating of the photograph alone as a covariate adjusted the error mean square for the combined rating by 4%, while treating the rating of the written information alone as a covariate adjusted the error mean square by 47%. Using both the rating of the photograph and the rating of the written information as covariates adjusted the error mean square by 49%. Thus, the written information alone had by far the most significant effect on the error mean square. The interactive effect of the photograph and written information when both were treated as covariates was negligible when compared to the effect of the written information alone.

The results of the analysis of variance and the various analyses of covariance were quite consistent. For all four analyses a statistically significant difference was observed between

TABLE 3

STANDARD NORMAL REGRESSION WEIGHTS AND MULTIPLE CORRELATIONS FOR APPLICANTS CONSTRUCTED OF CONTRASTING INFORMATION

Degree of favorability	Low intrarater stability		High intrarater stability	
	Low interrater agreement	High interrater agreement	Low interrater agreement	High interrater agreement
Unfavorable				
B_2	.11	.14	.13	.42
B_3	.44	.30	.59	.14
$R_{1.23}$	.46	.31	.60	.47
N	43	53	49	26
Favorable				
B_2	.18	.29	.11	.64
B_3	.79	.49	.81	.14
$R_{1.23}$	.84	.55	.84	.72
N	55	48	40	51

Note.—For abbreviations see note of Table 2.

TABLE 4

ANALYSIS-OF-VARIANCE AND ANALYSIS-OF-COVARIANCE RESULTS FOR APPLICANTS CONSTRUCTED OF COMPLEMENTARY INFORMATION

Source[a]	I. Anova				II. Anacova photographs				III. Anacova written				IV. Anacova photographs and written			
	df	*MS*	*F*	E²[b]	*df*	*MS*	*F*	E²	*df*	*MS*	*F*	E²	*df*	*MS*	*F*	E²
A	1	.11	.11	.00	1	.23	.20	.00	1	.30	.46	.00	1	.20	.32	.00
B	1	1412.83	1138.79**	.75	1	1233.62	1038.92**	.73	1	17.31	26.51**	.64	1	16.23	26.13**	.62
C	1	4.13	3.33	.01	1	5.51	4.64	.01	1	.00	—	.00	1	.06	.09	.00
AB	1	7.29	5.88		1	7.08	5.97		1	.17	.26		1	.19	.30	
AC	1	2.88	6 35		1	6.34	5.34		1	.45	.68		1	.27	.43	
BC	1	1.16	.94		1	5.38	4.53		1	1.40	2.14		1	.05	.08	
ABC	1	.07	.06		1	.16	.14		1	.23	.46		1	.40	.64	
Error	375	1.24			374	1.19			374	.65			373	.63		

[a] Treatments: A = Intrarater agreement; B = Favorability; C = Interrater agreement.
[b] Epsilon squared. Peters and Van Voorhis' (1940) statistic for the unbiased estimate of the correlation ratio.
* $p < .05$.
** $p < .01$.

the mean rating for applicants constructed of favorable as opposed to unfavorable items of information. (See the adjusted means in Tables 6 and 7.) The other information properties did not have a significant effect on the mean ratings. Thus, although the covariates substantially reduced the total variance in the rating of the combined information, a significant difference remained between the mean ratings for applicants constructed of favorable as opposed to unfavorable information. Apparently some meaningful variance exists in the rating of the photograph and written information combined beyond that contributed by the ratings of the written information alone and the photograph alone.

The relationship between the main treatment effects and the rating is indicated by epsilon squared in Table 4 (Peters & Van Voorhis, 1940). Epsilon squared is an unbiased estimate of the correlation ratio and, as such, indicates the strength of the association between a treatment effect and the dependent variable. In the absence of a covariate, epsilon squared was .75 between the favorability of information and the final rating. When the rating of the photograph was treated as a covariate, epsilon squared for the favorability of information dropped only slightly to .73. When the written information was treated similarly, epsilon squared dropped to .64, and when both the photograph and written information were treated as covariates, the favorability of the information still produced an epsilon squared of .62. Thus, not only was the mean difference statistically significant, but the degree of association between the favorability of information and the final rating was also significant.

Another way to show the strength of the relationship between the favorability of the information and the overall rating of the ap-

TABLE 5

ANALYSIS-OF-VARIANCE AND ANALYSIS-OF-COVARIANCE RESULTS FOR APPLICANTS CONSTRUCTED OF CONTRASTING INFORMATION

Source	I. Anova				II. Anacova photographs				III. Anacova written				IV. Anacova photographs and written			
	df	*MS*	*F*	E²	*df*	*MS*	*F*	E²	*df*	*MS*	*F*	E²	*df*	*MS*	*F*	E²
A	1	.73	.63	.00	1	.86	.76	.00	1	.32	.44	.00	1	.24	.35	.00
B	1	1009.43	872.20**	.71	1	1021 28	913.20**	.72	1	.86	1.19	.00	1	1.63	2.36	.00
C	1	15.55	13.40**	.03	1	17 02	15.22**	.04	1	1.20	1.66	.00	1	1.64	2.38	.00
AB	1	16.76	14.48**		1	15.14	13.54**		1	1.32	1.83		1	1.02	1.48	
AC	1	17.06	14.74**		1	19.08	17.06**		1	2.95	4.09		1	3.81	5.51	
BC	1	6.31	5.43*		1	9.03	8.08**		1	.03	.04		1	.09	.13	
ABC	1	.87	.75		1	1.64	1.47		1	.16	.22		1	.51	.74	
Error	357	1.16			356	1.12			356	.75			355	.68		

Note.—See footnotes of Table 4.
* $p < .05$.
** $p < .01$.

atic effect beyond that contributed by the best linear combination of the separate elements of information. However, when the information was complementary, there was a significant and systematic effect due to the favorability of the information used to construct the hypothetical applicants. This effect was greater than the best linear combination of the separate pieces of information.

REFERENCES

BOLSTER, B. I., & SPRINGBETT, B. M. The reaction of interviewers to favorable and unfavorable information. *Journal of Applied Psychology*, 1961, 45, 97–103.

BROWN, B. W., JR., LUCERO, R. J., & FOSS, D. B. A situation where the Pearson correlation coefficient leads to erroneous assessment of reliability. *Journal of Clinical Psychology*, 1962, 18, 95–97.

CARLSON, R. E. Selection interview decisions: The effect of interviewer experience, relative quota situation and applicant sample on interviewer decisions. *Personnel Psychology*, 1967, in press. (a)

CARLSON, R. E. Selection interview decisions: The effect of mode of applicant presentation on some outcome measures. *Personnel Psychology*, 1967, in press. (b)

CARLSON, R. E., & MAYFIELD, E. C. Selection interview decisions: The effect of type of information on inter- and intra-interviewer agreement. *Personnel Psychology*, 1967, in press.

HAYS, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.

MAYFIELD, E. C. The selection interview—a re-evaluation of published research. *Personnel Psychology*, 1964, 17, 239–260.

MAYFIELD, E. C., & CARLSON, R. E. Selection interview decisions: First results from a long-term research project. *Personnel Psychology*, 1966, 18, 41–55.

PETERS, C. C., & VAN VOORHIS, W. R., *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.

SPRINGBETT, B. M. Factors affecting the final decision in the employment interview. *Canadian Journal of Psychology*, 1958, 12, 13–22.

WEBSTER, E. C. *Decision making in the employment interview*. Montreal: Industrial Relations Centre, McGill University, 1964.

(Early publication received June 19, 1967)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 469–475

JOB ANALYSIS BY MULTIDIMENSIONAL SCALING

KENNETH R. BROWN[1]

United States Civil Service Commission

Multidimensional scaling methods were used to determine the dimensions of interpersonal relations in a specific job setting. 18 behavior statements relating to interpersonal relations in a management-analyst position in the Federal government were developed. Job incumbents judged the similarity of the statements. Data were collected and analyzed by both the traditional multidimensional scaling method and the A-technique. Results indicated the dimensions of interpersonal relations in the job. A comparison of the 2 different multidimensional approaches indicated that they produced similar results. In view of certain administrative advantages of the A-technique, further use of the A-technique in analyzing job domains seems justified.

An analysis of job content or worker activity is a basic part of many types of personnel research investigations. For example, some form of job or worker analysis seems essential in the areas of test construction and validation, training development and evaluation, personnel appraisal development, etc. Information resulting from such an analysis frequently is used to direct the approach taken in subsequent research and development tasks and to provide a basis for directly relating the measuring instruments, training courses, etc., to the job with which they will be used.

Quantitative job analyses frequently are based on evaluations made in accordance with some a priori definition selected by the investigator; that is, judges are instructed to evaluate specified aspects of the job in terms of a particular rating consideration. In such procedures, the investigator assumes that he has selected the quality most appropriately related to the decision. In many instances, however, raters' true judgments may not represent a single dimension but may be the result of a composite which includes several dimensions. A technique which enables the rater to evaluate the work without restrictions on the way in which he perceives it would appear to be a better approach to derive the actual factors of the job.

A similar situation occurs when considering the nature of each of the separate factors identified as essential in the job. If, for example, administrative ability is determined to be involved in a certain occupation, a question arises as to what the term administrative ability actually means in this job context, that is, what are the dimensions underlying the concept. General terms such as human relations, administrative ability, etc., are rather vague, with different meanings associated with them by different people. In order to identify the specific job aspects which should be considered in establishing appropriate qualification requirements and in developing devices for obtaining valid predictions of successful job performance, an analysis of each important factor should be conducted in such a way as to reveal the basic dimensions underlying the factor.

Multidimensional scaling analysis (Torgerson, 1958) provides a method for determining the fundamental dimensions of complex variables for which the psychological continuum is composed of a complex of two or more dimensions. This method differs from traditional unidimensional scaling in that, with multidimensional scaling, the minimum dimensionality of a set of stimuli is determined without influencing the raters' judgment by requiring them to evaluate the stimuli in terms of the investigator's arbitrarily chosen definitions. The idea of multidimensional scaling originated with Richardson (1938) but did not receive much attention until research by Torgerson (1952), Abelson (1954), Messick (1956b), and others illustrated its potential applications to various psychological problems.

[1] This article is based on a doctoral dissertation completed at the University of Maryland under the direction of C. J. Bartlett.

The traditional multidimensional scaling technique consists of several stages. First, a set of scale values representing relative distances between stimuli is obtained. Then, these relative distances are converted to absolute interstimulus distances by the use of an additive constant. Finally, the dimensionality of the space that describes the set of absolute distances is determined, and the projections of the stimuli on the axes of this space are obtained. Schultz and Siegal (1964) and Siegal and Pfeiffer (1965) recently applied the traditional multidimensional scaling technique in a job-analysis situation for the purpose of developing a criterion measure for the position of aviation electronics technician in the Navy. The stimuli for these studies covered job tasks ranging from the performance of housekeeping duties and routine operations to the use of schematics and test equipment.

Andrews and Ray (1957) have suggested a different multidimensional approach which involves the transformation of the proportions of common elements between stimuli into estimates of correlation coefficients which can then be factored by the usual factor-analytic methods. This multidimensional procedure was named the A-technique. The method is based on the logic that an acceptable estimate of the proportion of common elements in sets of data is the coefficient of determination (r^2), and, therefore, it is equally true that an acceptable estimate of r is the square root of the proportion of common elements.

The present research was an investigation of the use of multidimensional scaling to study an important job factor or domain in depth in order to understand its meaning and underlying dimensions as perceived by individuals currently working in the position. The job domain of interpersonal relations was selected for study, and both the traditional multidimensional scaling method and the A-technique were applied. The objectives of the research were: (*a*) to demonstrate the use of the multidimensional scaling method in the analysis of job domains, (*b*) to compare the results of two multidimensional approaches, and (*c*) to determine by multidimensional analysis the dimensions underlying the concept of interpersonal relations in a specific job setting.

METHOD

The investigator, with the aid of an occupational specialist, developed 53 behavior statements that were related to the domain of interpersonal relations for the position of management analyst in the Federal government. Six management analysts, working individually, divided the 53 statements into 3 groups by selecting the 18 items which described behavior considered most important in the management-analyst position and the 18 items considered least important. The remaining items were considered to be of intermediate importance. The statements were scored on a scale of 1 (for most important) to 3, depending on the group into which they were placed. The mean rating for each behavior statement was calculated, and the 18 statements with the lowest mean ratings were selected for inclusion in the study. These 18 statements are:

1. Presents ideas and recommendations so that affected persons or groups are not offended.
2. Cooperates with fellow workers to get a job done.
3. Maintains good relations with supervisor and other higher officials.
4. Expresses himself effectively—to the point.
5. Deals effectively with members of other organizations.
6. Helps people who ask for assistance.
7. Promotes the use of new ways of doing things.
8. Interviews supervisors and employees to collect information.
9. Convinces officials to use management-analysis techniques.
10. Negotiates with management for approval of proposals and recommendations.
11. Explains results and recommendations of studies to management.
12. Develops an acceptable course of action when different points of view are advocated by different people.
13. Coordinates plans with all others involved.
14. Listens to viewpoints of others.
15. Accepts criticism without getting resentful.
16. Works as a member of a team to study a management problem.
17. Makes himself accessible to others.
18. Sells changes in management practices when operating officials prefer previous methods or have "pet" ideas.

The subjects (*S*s) rated the 18 stimuli by three different methods:

1. In terms of the multidimensional method of successive intervals as suggested by Messick (1956a). This method required *S*s to rate the similarity of all possible pairs of stimuli. Each pair was rated on a 7-point scale.
2. In terms of the method of nonserial matching, Case III of the A-technique. In this method, the

stimuli were presented to *S* in the form of two lists, each containing the same behavior statements but in different order. For each statement in the first list, *S* selected the statement from the second list which was most similar to it.

3. In terms of four specified rating considerations (Consideration, Directing and Controlling, Difficulty, and Importance) which were believed to be pertinent to the concept being studied. Each stimulus was rated on a 7-point scale in accordance with each of the four rating considerations.

Information obtained by the third rating method was used as an aid in interpreting factors resulting from analysis of the data produced by the first two rating methods.

The sample consisted of 120 civilian employees in the position of management analyst in an agency of the Federal government. Data collected from the sample by application of the multidimensional method of successive intervals were analyzed by the traditional multidimensional scaling method, and data collected by application of the method of nonserial matching were analyzed by the A-technique.

In the traditional method, the successive-intervals solution developed by Diederich, Messick, and Tucker (1957) was used to obtain the matrix of relative interstimulus distances. The procedure suggested by Messick and Abelson (1956) was used to determine the additive constant for use in translating the scale values to absolute distances measured from an absolute zero point. This method involves an initial estimation of the constant and then an iteration procedure to eliminate the distortion introduced by the estimation. The resulting matrix of absolute interstimulus distances, with zeros in the diagonals, was transformed to a scalar-product matrix by the procedure suggested by Torgerson (1952) who reasoned that the origin should be placed at the centroid of all the stimuli.

The scalar-product matrix was factor analyzed by the principal-axes method, and the major factors were rotated by the Varimax method suggested by Kaiser (1958). The factor loadings, in this case, may legitimately exceed 1.00 since the values are not cosines of angles but are scale values of the projections of the stimuli on the axes of the space.

In the A-technique, the frequency of each pairing of stimuli was plotted, and the corresponding proportions (ratio of frequency of selection to possible frequency of selection) were computed. A square-root transformation of these proportions was made to produce estimates of the correlations between stimuli. The resulting matrix was factor analyzed by the principal-axes method, and rotation was accomplished by the Varimax method.

RESULTS

One of the objectives of the present investigation was to compare the A-technique with the traditional multidimensional scaling method to determine if these two approaches lead to similar results. As the first comparative analysis, the 20 largest and 20 smallest absolute interstimulus distances from the traditional multidimensional scaling analysis were compared with the estimated correlation coefficients derived from the transformation procedure used in the A-technique. The results are contained in Table 1.

Since small absolute interstimulus distances indicate that the stimuli are perceived as being closely related, one would expect that, if the stimuli were perceived in the same manner in each of the multidimensional analyses, stimulus pairs with small absolute distances also would have substantial correlation coefficients. Similarly, stimulus pairs with large absolute interstimulus distances would have produced small correlation coefficients in the A-technique analysis. The results shown in Table 1 substantiate this expectation. Of the correlation coefficients for the 20 largest interstimulus distances, 16 were .00. The highest correlation was only .11. On the other hand, the correlation coefficients for the 20 smallest absolute distances ranged from .25 to .83, with 15 of the coefficients being above .40.

The traditional multidimensional analysis resulted in the identification of five major

TABLE 1

COMPARISON OF THE 20 LARGEST AND 20 SMALLEST ABSOLUTE INTERSTIMULUS DISTANCES WITH THEIR CORRESPONDING ESTIMATED CORRELATION COEFFICIENTS

Smallest distances			Largest distances		
Stimulus pair[a]	Absolute distance	Estimated correlation	Stimulus pair[a]	Absolute distance	Estimated correlation
6–17	.27	.83	8–18	3.36	.00
2–16	.52	.80	6–15	3.35	.00
7–18	.75	.60	9–17	3.21	.00
4–11	.77	.51	6–18	3.19	.00
8–14	.82	.64	7–15	3.17	.00
5–13	.91	.48	9–15	3.15	.00
10–11	.93	.70	6–10	3.15	.00
7–9	.93	.49	17–18	3.13	.00
1–12	1.01	.56	8–15	3.12	.00
2–13	1.03	.43	1–8	3.12	.06
10–18	1.06	.34	8–11	3.11	.00
14–17	1.08	.36	8–10	3.06	.06
12–13	1.11	.43	4–15	3.05	.00
1–5	1.14	.47	1–17	3.02	.00
1–18	1.14	.44	2–9	2.99	.00
9–18	1.20	.50	15–17	2.97	.11
3–5	1.21	.41	2–11	2.96	.00
5–12	1.23	.39	11–17	2.95	.06
5–14	1.25	.25	4–14	2.94	.00
4–18	1.26	.26	8–9	2.91	.00

[a] Numbers refer to the stimuli listed under the Method section.

TABLE 2

ROTATED FACTOR MATRIX FOR THE TRADITIONAL MULTIDIMENSIONAL SCALING ANALYSIS

Stimulus[a]	Factor				
	I	II	III	IV	V
1	.64	−.28	−.65	.22	.95
2	−.70	−.21	.17	−1.20	−.63
3	−.12	1.06	.04	.27	.35
4	.18	−.46	.18	1.26	.16
5	−.07	.14	.05	−.12	.63
6	−1.91	−.30	−.10	−.20	.09
7	.25	−1.03	−.54	.56	−.56
8	−.23	.04	2.11	− .12	−.18
9	.46	−.26	−.20	1.42	−.30
10	.87	.02	−.58	.55	−.07
11	.25	−.04	−.59	1.21	.01
12	.84	−.28	−.25	−.53	.46
13	.09	−.41	.00	−.83	.46
14	−.21	.67	.65	−1.02	.05
15	.52	1.88	−.23	−.69	−.55
16	−.12	−.48	.65	−.85	−.89
17	−1.75	.14	.20	−.62	.00
18	1.01	−.20	−.88	.69	.01

[a] Numbers refer to the stimuli listed under the Method section.

factors. The rotated factor matrix is found in Table 2.

The loadings suggested the following factor descriptions:

Factor I—Tact in Personal Contacts. Major Items: Sells changes in management practices when operating officials prefer previous methods or have "pet" ideas, negotiates with management for approval of proposals and recommendations, develops an acceptable course of action when different points of view are advocated by different people, and presents ideas and recommendations so that affected persons or groups are not offended.

Factor II—Accepting Others' Viewpoints. Major Items: Accepts criticism without getting resentful, maintains good relations with supervisor and other higher officials, and listens to viewpoints of others.

Factor III—Information Collection. Major Items: Interviews supervisors and employees to collect information, listens to viewpoints of others, and works as a member of a team to study a management problem.

Factor IV—Verbal Persuasiveness. Major Items: Convinces officials to use management-analysis techniques, expresses himself effectively—to the point, explains results and recommendations of studies to management, and sells changes in management practices when operating officials prefer previous methods or have "pet" ideas.

Factor V—Controlling People through Consideration. Major Items: Presents ideas and recommendations so that affected persons or groups are not offended, deals effectively with members of other organizations, develops an acceptable course of action when different points of view are advocated by different people, and coordinates plans with all others involved.

The A-technique analysis resulted in the identification of seven factors. The rotated factor matrix is found in Table 3.

The loadings resulting from the A-technique analysis suggested the following factor descriptions:

Factor I—Controlling People through Consideration. Major Items: Deals effectively with members of other organizations, presents ideas and recommendations so that affected persons or groups are not offended, develops

TABLE 3

ROTATED FACTOR MATRIX FOR THE A-TECHNIQUE ANALYSIS

Stimulus[a]	Factor						
	I	II	III	IV	V	VI	VII
1	.68	.06	−.02	−.01	.26	.14	−.02
2	.17	.11	.19	.87	.02	−.04	−.01
3	.41	.36	.04	.17	.02	.10	−.02
4	.26	−.15	−.01	−.12	.21	.49	.26
5	.77	.01	.13	.03	−.04	.10	.24
6	.08	−.04	.93	.12	.06	−.03	.02
7	.02	−.04	.06	.10	.81	.08	.05
8	.08	.04	.04	.17	.03	.03	.90
9	.07	−.02	−.01	.01	.62	.23	.03
10	.17	.09	−.02	.02	.28	.71	−.07
11	.05	.02	.02	.05	.05	.95	−.02
12	.65	.12	.02	.05	.21	.07	−.02
13	.54	−.02	.04	.42	−.04	.05	.03
14	.11	.70	.19	.06	.00	−.05	.62
15	.09	.92	.01	.04	.02	.00	.00
16	.04	.03	−.01	.90	.04	.03	.20
17	.05	.14	.93	.05	−.03	.03	.08
18	.28	.11	−.03	−.08	.80	.07	−.03

[a] Numbers refer to the stimuli listed under the Method section.

an acceptable course of action when different points of view are advocated by different people, and coordinates plans with all others involved.

Factor II—Accepting Others' Viewpoints. Major Items: Accepts criticism without getting resentful, listens to viewpoints of others, and maintains good relations with supervisor and other higher officials.

Factor III—A doublet: Helps people who ask for assistance and makes himself accessible to others. These items reflect situations in which other people contact the management analyst to request his services; a tentative name for this potential factor might be Providing Requested Services.

Factor IV—Cooperation. Major Items: Works as a member of a team to study a management problem, cooperates with fellow workers to get a job done, and coordinates plans with all others involved.

Factor V—Persuasion through Logic. Major Items: Promotes the use of new ways of doing things, sells changes in management practices when operating officials prefer previous methods or have "pet" ideas, and convinces officials to use management-analysis techniques.

Factor VI—Verbal Expression. Major Items: Explains results and recommendations of studies to management, negotiates with management for approval of proposals and recommendations, and expresses himself effectively—to the point.

Factor VII—Information Collection. Major Items: Interviews supervisors and employees to collect information, listens to viewpoints of others, and expresses himself effectively—to the point.

A comparison was made of the factors resulting from each multidimensional analysis. The comparison is summarized in Table 4. The factors which appear to be most closely related are found in the same row of the table. In addition, the product-moment correlation between the factor loadings for the two related factors is shown in the last column of the table. Since the traditional multidimensional method produces bipolar factors while the A-technique does not, one would not expect congruence coefficients to be especially high. This condition would be the result of the A-technique procedure which restricts the interstimulus correlation matrix to zero or positive correlation coefficients. Therefore, the product-moment correlation or shape coefficient was used to compare factors in this study.

TABLE 4

COMPARISON OF FACTORS OBTAINED FROM THE A-TECHNIQUE WITH FACTORS OBTAINED FROM THE TRADITIONAL MULTIDIMENSIONAL SCALING METHOD

Traditional multidimensional scaling factor	A-technique factor	r[a]
I. Tact in personal contacts	III. Providing requested services	−.90
II. Accepting others' viewpoints	II. Accepting others' viewpoints	.88
III. Information collection	VII. Information collection	.88
IV. Verbal persuasiveness	IV. Cooperation	−.59
	V. Persuasion through logic	.59
	VI. Verbal expression	.68
V. Controlling people through consideration	I. Controlling people through consideration	.81

[a] Product-moment correlation between loadings on the two related factors. $N = 18$.

DISCUSSION

Previous research studies (e.g., Halpin & Winer, 1957; Rupe, 1951) conducted for the purpose of determining the dimensions of interpersonal behavior in a work situation generally have resulted in the identification of two major factors: sensitivity to the feelings of other people, and directing or controlling the behavior of other people. These dimensions have been referred to by various names, perhaps the most popular of which have been Consideration and Initiating Structure. The findings of the present research, while specifically related to the position of management analyst, also are related to these major factors. For example, Factor IV from the traditional multidimensional analysis and Factor V from the A-technique analysis were interpreted as "persuasion" factors which seem related to the factor of Initiating Structure.

Factor V of the traditional analysis and its counterpart factor from the A-technique, Factor I, are of particular interest in this regard. These two factors are related to the dimension of Consideration as shown by the correlation of stimulus factor loadings with the ratings assigned to the stimuli when evaluated by the sample in terms of the

degree of consideration involved in the behavior described by the statements. The correlations for these two factors were .62 and .66, respectively. The two factors also correlated .63 and .59, respectively, with the ratings assigned in terms of the degree to which the behavior involves the ability to direct or control the behavior of other people. The correlation of these factors with both Consideration and Directing and Controlling may be a finding related to staff positions, such as management analyst, in which direct authority for implementation of proposals is vested in line officials. In many instances, the staff man gains the acceptance of line officials by taking actions such as compromising to reach an agreement, cooperating with them so that they cooperate with him, etc. In other words, he must be considerate of their opinions, feelings, etc., so that he can maintain some degree of control over their behavior. Some recent research conducted by other investigators (Fleishman & Harris, 1962; Fleishman & Peters, 1962; Rim, 1965) also has suggested that the individuals who are most effective in influencing others are those rated high on both Consideration and Initiating Structure.

The results shown in Table 4 indicate considerable agreement in the factors identified in the two different multidimensional analyses. Factors I, II, III, and V from the traditional multidimensional scaling analysis each appear to have a counterpart factor in the A-technique analysis. The investigator's interpretation of Factor I from the traditional analysis, however, varies considerably from the interpretation of its counterpart factor, Factor III, from the A-technique analysis. The negative pole of the traditional multidimensional scaling factor definitely coincides with the single pole of the A-technique factor which was tentatively interpreted as Providing Requested Services. However, the positive pole of the traditional multidimensional scaling factor, especially the inclusion of items such as 1, 12, and 15, suggested that the factor involved tact. The tentative interpretation of the A-technique factor was based on a doublet consisting of Items 6 and 17 which did not permit an interpretation involving tact.

A second major difference between the two sets of factors is that the traditional method produced five factors while the A-technique produced seven factors. The additional A-technique factors seem to occur in the form of a finer division of the Verbal Persuasiveness factor (Factor IV) from the traditional multidimensional analysis. Andrews and Ray (1957) have indicated that the A-technique sometimes produces two individual factors rather than a single bipolar factor because negative roots are not used in the square-root transformation which is applied in the A-technique procedure for estimating the interstimulus correlation coefficients. The negative pole of the Verbal Persuasiveness factor is composed of "cooperation items" as is Factor IV from the A-technique analysis. The positive pole of the Verbal Persuasiveness factor includes the items which involve persuasion as well as those which involve primarily verbal expression. The A-technique, on the other hand, makes a distinction between these two types of items. Thus Factor IV from the traditional analysis appears to be a composite of Factors IV, V, and VI resulting from the analysis by the A-technique.

The differences in factors identified by the two multidimensional techniques also point up the problem of the number of factors to be extracted. In the present research, the number of factors rotated in the A-technique analysis was based on the number of factors with eigenvalues greater than 1.00. In the traditional multidimensional approach, the decision as to the rank of the matrix is somewhat more subjective. Normally, the investigator attempts to determine the minimum dimensionality that provides an adequate fit of the data. In the present research, this approach led to the decision to retain five factors. However, such a decision is not necessarily equated to that which was employed with the A-technique. If six factors are rotated for the traditional multidimensional analysis, the sixth factor is a verbal factor with high loadings on Items 4 and 11, as is the case with Factor VI of the A-technique analysis. Thus, the retention of six factors in the traditional analysis brings the results even more in line with those obtained with the A-technique.

TABLE 5

COMPARISON OF FACTORS FROM THE TWO MULTIDIMENSIONAL ANALYSES WHEN SIX FACTORS ARE ROTATED IN THE TRADITIONAL ANALYSIS

Traditional multidimensional scaling factor	A-technique factor	r[a]
I. Tact in personal contacts	III. Providing requested services	−.91
II. Accepting others' viewpoints	II. Accepting others' viewpoints	.83
III. Information collection	VII. Information collection	.90
IV. Persuasion through logic	IV. Cooperation	−.62
	V. Persuasion through logic	.71
V. Controlling people through consideration	I. Controlling people through consideration	.71
VI. Verbal expression	VI. Verbal expression	.81

[a] Product-moment correlation between loadings on the two related factors. $N = 18$.

The correlations between factors are shown in Table 5.

The overlap in interpretation of factors resulting from the two separate multidimensional analyses and the indication of a relationship between the relative magnitudes of the stimulus loadings on counterpart factors indicate that the two different methods of multidimensional analysis produced similar results. The factors of interpersonal relations in the management-analyst position suggested by the results of the various multidimensional analyses are: Tact in Personal Contacts, Accepting Others' Viewpoints, Information Collection, Cooperation, Persuasion through Logic, Verbal Expression, and Controlling People through Consideration.

The A-technique (Case III) has several advantages over traditional multidimensional scaling in that the method of presentation of the data is relatively simple, and the task imposed on the subject is relatively easy to accomplish. The ease with which data can be collected for the A-technique permits the expansion of the list of stimuli to cover more job behaviors, thereby decreasing the possibility of omitting some significant job factors. In view of these administrative advantages, further use of the A-technique in lieu of traditional multidimensional scaling for the analysis of job domains seems justified.

REFERENCES

ABELSON, R. P. A technique and a model for multidimensional attitude scaling. *Public Opinion Quarterly*, 1954, **18**, 405–418.

ANDREWS, T. G., & RAY, W. S. Multidimensional psychophysics: A method of perceptual analysis. *Journal of Psychology*, 1957, **44**, 133–144.

DIEDERICH, G. W., MESSICK, S. J., & TUCKER, L. R. A general least squares solution for successive intervals. *Psychometrika*, 1957, **22**, 159–174.

FLEISHMAN, E. A., & HARRIS, E. F. Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology*, 1962, **15**, 43–56.

FLEISHMAN, E. A., & PETERS, D. R. Interpersonal values, leadership attitudes and managerial success. *Personnel Psychology*, 1962, **15**, 127–143.

HALPIN, A. W., & WINER, B. J. A factorial study of the leader behavior descriptions. In R. M. Stogdill & A. E. Coons (Eds.), *Leader behavior: Its description and measurement.* Bureau of Business Research Monograph 88, Ohio State University, 1957. Pp. 39–51.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, **23**, 187–200.

MESSICK, S. J. An empirical evaluation of multidimensional successive intervals. *Psychometrika*, 1956, **21**, 367–375. (a)

MESSICK, S. J. The perception of social attitudes. *Journal of Abnormal and Social Psychology*, 1956, **52**, 57–66. (b)

MESSICK, S. J., & ABELSON, R. P. The additive constant problem in multidimensional scaling. *Psychometrika*, 1956, **21**, 1–15.

RICHARDSON, M. W. Multidimensional psychophysics. *Psychological Bulletin*, 1938, **35**, 659–660.

RIM, Y. Leadership, attitudes and decisions involving risk. *Personnel Psychology*, 1965, **18**, 423–430.

RUPE, J. C. When workers rate the boss. *Personnel Psychology*, 1951, **4**, 271–290.

SCHULTZ, D. G., & SIEGAL, A. I. The analysis of job performance by multidimensional scaling techniques. *Journal of Applied Psychology*, 1964, **48**, 329–335.

SIEGAL, A. I., & PFEIFFER, M. G. Factorial congruence in criterion development. *Personnel Psychology*, 1965, **18**, 267–279.

TORGERSON, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, **17**, 401–419.

TORGERSON, W. S. *Theory and methods of scaling.* New York: Wiley, 1958.

(Received September 21, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 476–480

A MULTIDIMENSIONAL SCALING ANALYSIS OF THE JOB OF CIVIL DEFENSE DIRECTOR[1]

ROBERT J. SMITH[2] AND ARTHUR I. SIEGEL

Applied Psychological Services, Wayne, Pennsylvania

An examination was performed, through multidimensional analytic scaling techniques, of the complex job of the Office of Civil Defense (OCD) Director. 35 regional and Headquarters staff judged the similarity of 34 job-relevant task functions culled from OCD plans of 7 states. 3 bipolar factors emerged from the subsequently factored matrix: (1) internal vs. external system maintenance, (2) routine vs. emergency programming, and (3) resource use vs. resource evaluation. A 4th factor, labeled emergency system integration, was less clear and appeared unipolar. It is concluded that multidimensional scaling analysis is a practical approach for defining complex jobs. Such defining would permit subsequent unidimensional measurement. The factors found may be used for selection, training, etc., of OCD directors.

Multidimensional scaling analysis has been shown to be an extremely practical technique for the classification or ordering of complex entities. The method has been demonstrated to produce results which agree with those obtained through other techniques (Helm, 1959), and has been applied to areas ranging from studies of attitudes (Messick, 1954) to relatively simple job skills (Siegel & Schultz, 1964). The important job of Civil Defense Director within the Office of Civil Defense (OCD) offered a particular challenge to the present investigators because: (*a*) a plethora of skills was clearly involved; (*b*) reducing them to meaningful dimensions without altering their essential complexity would permit, among others, subsequent performance evaluation and an approach to selection and training; and (*c*) a clearer delineation of similar supervisory-level positions both in and out of government might be facilitated.

The relevance and rationale for the present multidimensional scaling procedure are detailed elsewhere (Siegel & Schultz, 1964) and will not be elaborated herein.

Method

As with any standard job analysis it was first necessary to develop a list of the tasks performed by Civil Defense Directors. To this end "task cards" describing single aspects of the job were developed. These single statements were couched in the language contained in the plans and publications of OCDs of seven states having wide geographic coverage.[3]

The OCD publications of these states were analyzed and a library of approximately 10,000 task data cards was established. These data cards were reduced to approximately 1,500 tasks cards through generalizing and summarizing to a common language suitable to the publications of all seven states. Further summary and generalization yielded 300 statements related to state-level OCD operations. These 300 statements were classified according to several schemes, including relating the action to the appropriate service or position within the civil defense organization.

For purposes of the present study, only tasks relating to the civil defense agency, its director, and deputy director, were considered. These items were reconsidered, combined, rephrased, and recast to produce a list of 34 task functions, which retained as much of the sense of the original items as possible and which were associated with the position of the OCD director (Table 1).

Stimuli

The Civil Defense Task Inventory was constructed in accordance with the method of equal-appearing intervals. The stimuli to be judged were presented in booklet format. At the top of each page of the booklet, 1 of the 34 task functions was shown. Below it, at the left side of the page, the functions appeared which were to be compared with the

[1] This study was completed under Contract OCD-PS-64-30 between Applied Psychological Services and the Office of Civil Defense, Department of the Army, while the senior author was a member of the staff of Applied Psychological Services. The authors express their indebtedness to Dorothy Fletcher, D. MacPherson, and D. R. Saunders for their assistance.

[2] Now lecturer, University College, University of Maryland.

[3] States included were California, Missouri, North Carolina, Oklahoma, Oregon, Pennsylvania, and Vermont.

"standard" stimulus at the top. Within the booklet the pages were arranged (to control for sequence effects) in one of four orders. Each order was determined randomly and each booklet contained 20 pages. Item orders on any inventory page were randomly assigned. Only unidirectional comparisons were involved. Thus Page "A" contained 34 statements, 1 of which was enclosed in a box at the top of the page as a stimulus to be compared with the remaining 33. The initially boxed statement was then eliminated and, on Page "B," 1 of the remaining 33 was judged against 32 statements, and so on until only 2 items remained for comparison with one another.

Subject-Raters

The raters were 34 males and 1 female serving in a variety of supervisory-level OCD positions. Ten were from OCD Region II headquarters, 12 from OCD Region I headquarters, and 13 from OCD headquarters in Washington, D.C. Subjects (*Ss*) were screened on the basis of depth of experience and familiarity with the various aspects of OCD functioning. For example, their positions ranged from regional field officers to public information to planning specialists.

Procedure

The raters were to estimate the similarity of each listed function of the inventory with the single function at the top of each page. Judgments between this comparison item and each remaining item were made on an 11-point scale ranging from "very different" (11), through an undefined neutral point (6), to "very similar" (1). The rater's task was to make a check on the line under the number which corresponded to his conception of the similarity of the given "standard" stimulus (boxed) task function and the comparison task function. The *S*, hence, recorded (i.e., assigned a number to) his own response.

The data were collected separately at each location. Prior to completing the form, the judges were given a brief explanation of the overall purposes of the program. They were then asked to read the cover page of the inventory and to ask questions if necessary. The judges were then left free to proceed through the inventory at their own pace. Completion time among *Ss* ranged from 45 to 90 min.

TABLE 1

FINAL LIST OF THE OCD DIRECTOR'S TASK FUNCTIONS

1. Representing the executive branch in CD matters.
2. Organizing headquarters staff.
3. Assisting subordinate groups in establishing their CD organizations.
4. Directing subordinate CD groups.
5. Appointing CD technical advisory committees.
6. Assigning CD functions to other governmental and private agencies.
7. Administering the protected facilities program.
8. Prescribing mobilization procedures.
9. Alerting and mobilizing the CD system.
10. Assuring the operational readiness of the CD system.
11. Preparing and presenting a CD budget.
12. Preparing and implementing mutual aid agreements with related organizations.
13. Inspecting and reporting on installations assigned to or related to CD.
14. Assisting in legal actions arising from CD activities.
15. Issuing necessary emergency orders and instructions to the public.
16. Advising on needed CD legislation.
17. Prescribing information channels.
18. Relaying information from higher-level CD organizations.
19. Informing and advising executive branch and other governmental organizations on CD matters.
20. Assuring the proficiency of CD workers.
21. Programming the continuity of government.
22. Establishing the order of succession within the CD system.
23. Assuring the preservation of essential CD records.
24. Coordinating all CD activities.
25. Accounting for CD funds and property.
26. Arranging for fiscal and logistical support for CD activities.
27. Maintaining liaison with federal and state military groups.
28. Informing public and private groups of CD activities.
29. Conducting required research.
30. Evaluating potential emergencies.
31. Determining the availability of human and material resources.
32. Preparing CD plans.
33. Coordinating plans with those of other CD levels.
34. Conducting and evaluating CD tests.

RESULTS AND DISCUSSION

Multidimensional Scaling Analysis

For each function pair, the similarity judgments of the 35 *Ss* were pooled and median values obtained. For purposes of the present multidimensional scaling method, this scale value is interpreted as the relative psychological distance between the pairs of task functions. The goal of the analysis is to determine the number of axes in the space which account for the structure defined by these distances among all the stimuli, and to determine the projections of each stimulus on the axes.

Having attained the scaled between-task distance values, the Messick-Abelson (1956) general solution to the additive constant problem was applied to the data. This resulted in a value to be added to the original scaled dis-

negative and positive loadings involve tasks that represent essentially dichotomous functions when viewed with regard to the broadly defined labels assigned. Viewed as a whole, the factor names appear consistent with the description of the OCD director's major responsibilities and duties. One of the purposes of the present study was to investigate the utility of the multidimensional scaling analytic model for obtaining a parsimonious description of an administrative job, with specific reference to the job of the OCD director. The obtained results support a contention that the model appears applicable in these circumstances. The contention is supported on the basis of both the qualitative and the quantitative data.

From the qualitative point of view it is evident that the factors appear meaningful. "Meaningfulness" is probably a principal criterion to employ when a statistical-mathematical model is superimposed on sensory or qualitative attributes. Likewise, the factors are parsimonious. The 34 task functions, representing a complex psychological space, were reduced to four orthogonal factors which now may be more easily defined operationally.

From the quantitative point of view one may look to the "neatness" of the factor-analytic results. With respect to the factor matrix, there was a clean break after Factor 4. Moreover, the symmetry of the unrotated matrix, after the break, was quite acceptable. In addition the factors appear more or less consistent with the building system–operating system inherent in the OCD developmental cycle. Specifically, Factor 2 seemed to split along the lines of routine planning and emergency planning. Similarly, while Factor 1 relates to building-system maintenance, Factor 4 seemed concerned with operating-system maintenance. Finally, the four extracted factors accounted for 62% of the variance in the matrix.

CONCLUSIONS

The obtained results not only form a basis for describing the job functions of the OCD director, but can serve a number of other purposes. The factors may now be employed as a basis for developing unidimensional scales which can be used as the basis for performance evaluation (cf. Schultz & Siegel, 1964). Having determined, for example, that "resource evaluation" is an important dimension of the director's duties, performance measures relating to an individual's actual performance can be constructed.

Similarly, selection measures and training programs may be built along the extracted dimensions. Thus, the personnel subsystem may be developed in an integrated fashion, each aspect of which is based on a common and firm foundation. Each element of the system would be based on common dimensions and be logically interrelated. Moreover, the task functions involved in each factor may now be psychometrically scaled along some dimension such as importance for achieving the overall OCD mission. The development of a series of Factor X importance matrices might form the basis for an initial step in the development of a descriptive model of the director's job. Such a model could serve an important function in the development of performance evaluative measures.

REFERENCES

HELM, C. E. *A multidimensional ratio scaling analysis of color relations.* Princeton, N. J.: Princeton University and Educational Testing Services, 1959.

MESSICK, S. J. *The perception of attitude relationships: A multidimensional scaling approach to the structuring of social attitudes.* Unpublished doctoral dissertation, Princeton University, 1954.

MESSICK, S. J., & ABELSON, R. P. The additive constant problem in multidimensional scaling. *Psychometrika,* 1956, **21**, 1–15.

SAUNDERS, D. R. A computer program to find the best-fitting orthogonal factors for a given hypothesis. *Psychometrika,* 1960, **25**, 207–210.

SCHULTZ, D. G., & SIEGEL, A. I. Post-training performance criterion development and application: The development of unidimensional scales from the dimensions derived from a multidimensional scaling analysis of the job of the naval aviation electronics technician. Technical Report, 1964, Applied Psychological Services, Wayne, Pa., Contract Nonr-2279(00).

SIEGEL, A. I., & SCHULTZ, D. G. Post-training performance criterion development and application: A comparative multidimensional scaling analysis of the tasks performed by naval aviation electronic technicians at two job levels. Technical Report, 1964, Applied Psychological Services, Wayne, Pa., Contract Nonr-2279(00).

(Received October 10, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 481–490

UNDERLYING DIMENSIONS OF PERSONAL BACKGROUND DATA AND THEIR RELATIONSHIP TO OCCUPATIONAL CLASSIFICATION

MELANY E. BAEHR AND GLENN B. WILLIAMS

Industrial Relations Center, University of Chicago

In order to identify underlying dimensions of personal background data, 3 successive factor analyses were performed on the responses of a vocationally heterogeneous sample of 680 male Ss to a wide spectrum of commonly used personal-background-data items. Using the final factoring, an analysis of variance of scores derived from 15 interpretable 1st-order factors across 10 occupational groups showed significant F ratios ($p < .001$) for virtually all factors. A 2nd-order factor analysis yielded 5 uncorrelated factors, thought to represent broad behavior patterns associated with the needs and achievements of individuals. The study indicates relationships between the identified dimensions and occupational classification and provides a framework for future investigations of the dynamic relationships between biographical dimensions and occupational success.

Personal background data are often used in occupational selection and in vocational advisement and placement largely because this information is relatively easily obtained, is generally less subject to distortion than responses to personality questionnaires, and has good face validity since a great deal of research indicates that what a person has achieved or failed to achieve in the past will, with some degree of assurance, indicate what he may achieve in the future. Another advantage of these data is their contribution to multiple correlations of psychological test batteries against an empirical criterion since they are relatively independent of skill and ability measures which, themselves, generally show moderate to high intercorrelations.

When rigorous psychometric procedures are used to investigate the predictive validity of personal-history items, the original data are generally quantified and obtained from standard biographical questionnaires or application blanks. By far the most common approach has been to identify items which are valid predictors for specific criteria within a given occupational group. These items are then combined without weighting or are assigned differential weights in the calculation of a composite "score" for the questionnaire as a whole. Scores obtained from these procedures have been cross-validated for occupations ranging from seasonal employees (Dunnette & Maetzold, 1955), through production supervisors (Lockwood & Parsons, 1960), clerical workers (Guthrie, 1956; Minor, 1958), various sales personnel (Goldsmith, 1922; Mosel & Wade, 1951), to managerial personnel (Scollay, 1956) and research scientists and engineers (Buel, 1965; Smith, Albright, & Glennon, 1961). A considerable body of work has also been done in this area for personnel in the armed services (Elliot, 1960; Levine & Zachert, 1951; Seaquist, Barry, & Sells, 1956).

While it has been demonstrated that this usage of background data has often increased the predictive effectiveness of test batteries, it has been justifiably criticized on the basis of its empiricism and failure to relate obtained results to meaningful dimensions of behavior. This problem has been met by Levine and Zachert (1951) who subjectively classified items in terms of content. The validity of each item in the content category was then determined and if an acceptably large number of items showed significant relationships with the criterion, the content category was retained and the items differentially weighted. A more theoretical approach was taken by Super (1960) who postulated five major categories of experience which could be measured through biographical data.

In their study of research scientists, Morrison, Owens, Glennon, and Albright (1962)

met the criticism of empiricism by a factor analysis of life-history items which had been previously validated against three performance criteria. Their twin objectives were a better comprehension of the personal characteristics represented by the items and the use of the identified dimensions to examine the differential profiles of three criterion groups. Five factors were identified which accounted for 23% of the variance. The rather small proportion of variance accounted for, in the view of the authors, confirms the expectation that life-history data would reveal considerable uniqueness. While this "in-depth" study did generate factors which distinguished between different behavior patterns within the research-scientist occupational classification, the factors could not be expected to be of sufficient generality for use in the broader areas of counseling and placement or to contribute to an understanding of the dynamic relationships between personal background and general occupational success.

The approach followed in the present study was a factor analysis of the responses of a heterogeneous occupational sample to a wide variety of quantifiable personal-background-data items in multiple-choice format. The hope was that the resulting factors would be meaningful and significant background parameters for the general, male, employed population.

METHOD

Variables

A preliminary factor analysis was made of the responses of 242 male *Ss* in various industrial occupations to 80 rewritten background-data items derived from two principal sources, a patterned interview (McMurry, 1947) and a biographical questionnaire (Kerr, Baehr, & Burns, 1961). Thirteen factors were identified and interpreted. The present study grew directly from this research. New items were written which, hopefully, would contribute to some of the dimensions identified through the first factor analysis. In addition, the pool of items was expanded by recourse to the bank of such items developed and maintained by Division 14 of the American Psychological Association (Glennon, Albright, & Owens, 1963). This work resulted in the selection of 150 background-data items, expressed in a multiple-choice format in a research version of the Personal History Index (Baehr, Burns, & McMurry, 1963).

Subjects

A total of 680 male *Ss* responded to the research version of the questionnaire. The vocational heterogeneity of the sample may be gauged from the brief descriptions of the subgroups which follow.

Professionals ($n = 95$). The *Ss* consisted primarily of chemical engineers or chemists engaged either in product development or research. Most were in their late 30s, held a bachelor's degree in science or engineering, and earned between $11,000 and $15,000 a year.

Middle and Upper Executives ($n = 74$). These *Ss* ranged from third-level supervisors to top executives. Most were in their early or middle 40s, held a bachelor's degree in the arts, sciences, or business administration, and earned between $11,000 and $15,000 a year.

Junior Executives ($n = 26$). These *Ss* were largely engaged in administrative or higher-level clerical work. Most were in their middle 30s, held a bachelor's degree, and earned between $6,000 and $10,000 a year.

District Sales Managers ($n = 128$). This group consisted entirely of managers in the insurance business. Most held a bachelor's degree, were in their middle or late 30s, and earned approximately $15,000 a year.

Salesmen ($n = 128$). This subgroup consisted of insurance and paper-forms salesmen. The majority of both groups had from 1 to 4 yr. of college, and were in their early to middle 40s. Their annual salary ranges were, respectively, $6,000 to $10,000 and $11,000 to $15,000.

Foremen ($n = 51$). This subgroup consisted of first-line supervisors from metal, chemical, and petroleum industries. Most had completed high school, were in their middle to late 30s, and earned between $6,000 and $10,000 a year.

Hourly ($n = 79$). This subgroup included both skilled and semiskilled hourly personnel from engineering and manufacturing plants. Most of the skilled workers had completed high school while the semiskilled group averaged 9–11 yr. of schooling. The majority of both groups were in their early to middle 40s and earned between $6,000 and $10,000 a year.

Auditor Supervisors ($n = 61$). These *Ss* were classified by civil service as either GS-13 or GS-15. Most were between 45 and 50, held a bachelor's degree, and earned between $11,000 and $15,000 a year.

School Administrators ($n = 82$). This subgroup contained both lower-level school administrators and teachers aiming for an administrative position. The administrators held master's degrees, ranged in age from 30 to 34 yr., and earned between $6,000 and $10,000 a year.

Community School Directors ($n = 45$). This group consisted of directors in the community schools programs. The majority held bachelor's degrees and many held master's degrees in education or liberal arts. Ages ranged from 25 to 29 yr. and annual income from $6,000 to $10,000.

Procedure

The procedures described here were implemented through use of an IBM 7094 electronic computer.[1] The estimated means, variances, and kurtosis of the distribution of the responses of the total sample of 680 *Ss* were obtained for each of the 150 items. Item responses were dichotomized and the matrix of tetrachoric covariances calculated. Since the available factoring program would accept a maximum of 125 variables, a "goodness of variable program"[2] was used to identify the least promising variables. As a result, 26 variables were excluded from the first analysis. For the sake of completeness of the analysis of the data, however, these variables were included in two subsequent analyses as will be described in the Results section.

In each analysis, a principal axis solution was obtained with subsequent rotation of accepted factors to orthogonal simple structure using the equamax criterion (Saunders, 1962) and to oblique simple structure using the promax criterion (Hendriksen & White, 1964). The correlations between the oblique factor axes were obtained for the third and final factoring and a second-order analysis performed using the same factoring and rotational procedures that were employed in the first-order analyses.

The ability of the first-order oblique factors to discriminate between the occupational groups was investigated through an analysis of variance of both the calculated beta-weight and the unit-weight factor scores across the 10 occupational groups contributing to the sample. This was followed by a *t*-test analysis in which each occupational group was compared with every other for each of the factor scores.

Finally, the factors were examined according to the following four criteria: clearness of definition and of structure, number of items with significant loadings which defined the factor, the correlations between the respective beta-weight and unit-weight factor scores, and the Kuder-Richardson reliability coefficients, in order to make a tentative selection of factors with the most potential for operational use.

RESULTS

Factor Analyses

The successive first-order factor analyses provided a content analysis of the items with respect to duplication and relevance for the domain being studied. The first analysis, based on 124 items, yielded 22 factors. The majority of the rotated factors were interpretable but the structure was marred by an undesirably large number of items with low or insignificant factor loadings and the appearance of two doublets and a triplet. The latter were clearly due to duplication of items and were of no interpretative significance. Only one item from each of these factors was retained in subsequent analyses.

In all, 29 items were dropped from the second analysis because of duplication or insignificant loadings, and the 26 items which had not yet contributed to an analysis were substituted, making a total of 121 items. The second analysis yielded 21 factors. An examination of the oblique rotational solution showed that the major factors were relatively unchanged but there had been some sifting and sorting of items for factors which accounted for lesser percentages of variance. In general, the 26 items originally excluded on the basis of the "goodness of variable criterion" made very little contribution to the factorial structure.

In the third and final analysis, after the exclusion of all items on the basis of duplication, ambiguity, or nonrelevance, 105 items remained and were utilized. These items were composed of the 78 which had shown clearest factor loadings in either or both the previous analyses and an additional 27 items of rather doubtful status which were carried on the Dwyer (1937) extension procedure. These latter items did not affect the factorial structure but their loadings on the identified factors were calculated. Use of this procedure allowed the authors to be sure that potentially useful items were not excluded while maintaining clear factorial structure.

The final factoring yielded 15 factors which accounted for 66.6% of the correlation and 43.3% of the variance. In the oblique rotated matrix,[3] the items are identified by number, both for the 1963 edition and the revised version of the Personal History Index (Baehr, Burns, & McMurry, 1965).

The contribution of the items carried on the Dwyer extension procedure is minimal. Half of them show no loading greater than .20, and only two items have loadings as great as .30.

Interpretation of the First-Order Factors

The interpretations are given for the oblique rotational solution. The factors did not

[1] The authors are indebted to David R. Saunders for constructive criticism of the experimental design and for use of his computer programs.

[2] D. R. Saunders, personal communication, 1965.

[3] This table of results is available from the Industrial Relations Center, University of Chicago, upon request.

appear in the order given below but have been renumbered for convenience in presentation.

Factor 1 (*School Achievement*). The four highest loadings were on items which indicated that the school years were felt to be successful and stimulating, that the ranking in high school was in the top quarter, and that many school subjects were liked and few disliked. It was defined as: "Academic achievement, particularly in high school, but also, where applicable, at college. A general liking for, and adjustment to, the school environment."

Factor 2 (*Higher Educational Achievement*). The major loadings were on items which indicated that a bachelor's degree was held, that the respondent was 22 yr. or older before he got his first real job or one on which he was financially dependent, that he was 26 yr. or older before his first marriage or before his first child was born, and that his present annual earnings were over $10,000. The factor was defined as: "Special or technical accomplishment and qualification resulting in a relatively late vocational start and late assumption of family responsibilities. This pattern is characterized by eventual occupational and financial achievement."

Factor 3 (*Drive*). The five highest loadings were on items which indicated that the respondent had worked in three or more companies or had held three or more full-time jobs. He had supervised the work of others and had received merit raises on three or more past full-time jobs. He was 21 yr. or younger when first dependent on a job for financial support. The factor was defined as: "Inner drive to be outstanding in performance, to attain high goals even if this entails temporary setbacks, to supervise others, and to achieve success and advancement."

Factor 4 (*Leadership and Group Participation*). The major loadings were on items which indicated that the respondent had been active in three or more organizations during the last 3 yr. and had held one or more elective offices in the last 5 yr. He currently belonged to two or more state or national organizations. The minor loadings dealt with membership in honorary fraternities and participation in high school activities. The factor was defined as: "A desire to establish contact with others as shown by membership and interpersonal activity in organizations and an interest in influencing others through community and social activities. A high score suggests active participation and possible leadership in personal contact situations of various types."

Factor 5 (*Financial Responsibility*). Seven out of eight of the loadings were on items dealing with present financial status and handling of finances. The respondent indicated that present total assets, annual earnings, salary increases, and percentage of income saved were relatively high while monthly credit payments were less than 10% of income. He had owned stock in a company and carried $30,000 or more life insurance. The factor was defined as: "Ability to manage a personal economy of defined proportions—to earn, invest, save, and accumulate."

Factor 6 (*Early Family Responsibility*). The major loadings were on items which indicated that the respondent was 25 yr. or younger when he was first married and when his first child was born. He currently had two or more children, his wife did not earn, and he felt that his major accomplishment outside of work was in family activities. The factor was defined as: "Early marriage and establishment of a family, with the husband ordinarily being the sole provider. Demonstrated achievement in handling family's financial affairs. Outside the work situation, the greatest interest is in family activities."

Factor 7 (*Parental Family Adjustment*). The major loadings were on items which indicated that the respondent felt that, as a child, he had a happy home life and was often included in the leisure-time activities of the parents. He conferred with parents about occupational choice and, in his teens, usually went to his father or mother for help with a problem. The factor was defined as: "Development of realistic and constructive attitudes in the early family environment. This includes relationships between siblings, between parents, and between the child and the parents."

Factor 8 (*Situational Stability*). The four highest loadings were on items which indicated that the respondent was 40 yr. or older, had worked 10 yr. or more for one com-

pany, and did not think that his peak performance was some time in the future. He had lived 5 yr. or more at his last previous address. This factor was defined as: "Established security and stability in the work situation, resulting from past history of good performance. Presently more concerned with the maintenance of what has been achieved than with plans for improvement or development."

Factor 9 (*School Activities*). Major loadings were on items which indicated that the respondent participated in four or more high school activities, had three or more strong "outside" interests while in high school, and spent 10 or more hours a week on high school athletics. A minor loading indicated that he ranked, scholastically, in the top quarter of his high school class. The factor was defined as: "Major emphasis on active participation in athletic and extracurricular social activities at high school but also indication of good academic achievement."

Factor 10 (*Professional-Successful Parents*). The major loadings were on items which indicated that the father at least graduated from high school, and that his principal occupation was professional or managerial, or that he was self-employed, and that the respondent felt that the father was a successful parent. The factor was defined as: "A parental background characterized by a successful father, either self-employed or in one of the professions, and by material comfort and a happy home life."

Factor 11 (*Educational-Vocational Consistency*). This factor was essentially a triplet composed of three items dealing with the relationship between educational specialization and the job held, and with similarities in different full-time jobs that had been held by the respondent. It was defined as: "A preference for occupations which are highly related (or similar) and in line with educational interests and training."

Factor 12 (*Vocational Decisiveness*). This factor had only four sizable loadings, the three highest of which were on items which indicated that the respondent had not changed areas of specialization or vocational plans since high school. It was defined as: "Decisiveness in choosing an occupation, purposefulness in achieving the necessary qualifications, followed by an early start in the chosen occupation."

Factor 13 (*Vocational Satisfaction*). The major loadings were on items which indicate that the respondent would follow the same occupational route if he had it to do over again, that he would not change his area of specialization, and that his parents felt that he should choose the occupation which he liked best. There were a number of minor loadings which suggested satisfaction but also suggested rather limited goals. The factor was defined as: "Satisfaction with occupational choice and the expectation that peak performance will be some time in the future. There is, however, no evidence of consistency of application or of drive to achieve high standards of performance."

Factor 14 (*Selling Experience*). The three or four major loadings of this factor were on items which dealt with experience in buying and selling. There were a string of minor loadings on items which could be construed as being compatible with the background and activities of a salesman. However, the structure was not clear, and the factor, by definition, would discriminate between sales and other vocational groups. It was therefore not regarded as being of intrinsic interpretative significance. It was defined as: "Evidence of various kinds of selling experience, including door-to-door selling and transactions in real estate."

Factor 15 (*General Health*). All five items which loaded on this factor dealt with freedom from physical ailments and serious illnesses. It was defined as: "Generally better than average health over an extended period (child, adolescent, and adult). General freedom from physical ailments and from lost work time due to illness."

Interpretation of the Second-Order Factors

The correlations between the 15 primary factors are given in Table 1. The orthogonal and oblique rotational solutions for the factors derived from this matrix are given in Table 2. Loadings of .30 or greater for both the orthogonal (equamax) and oblique (promax) rotations are italicized in Table 2 and interpreted below. The factor patterns were

TABLE 1

CORRELATIONS BETWEEN THE 15 PRIMARY-FACTOR AXES

Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		.44	.26	.23	.33	.01	−.05	−.21	−.09	.10	.29	.19	.13	.08	−.11
2	.44		.18	.12	.27	.07	.07	−.20	−.01	.30	.25	.22	.17	.09	−.13
3	.26	.18		.23	.20	.20	−.08	.04	.05	.14	.06	.21	−.18	.28	−.25
4	.23	.12	.23		.07	.06	.11	−.36	.03	.21	.28	.51	.20	.09	.07
5	.33	.27	.20	.07		.37	−.04	.12	−.13	.24	.02	.14	.08	.08	−.01
6	.01	.07	.20	.06	.37		−.05	.01	.13	.20	−.05	.14	.07	.17	−.08
7	−.05	.07	−.08	.11	−.04	−.05		.05	.14	.20	−.01	.21	.22	−.09	.11
8	−.21	−.20	.04	−.36	.12	.01	.05		−.23	−.21	−.22	−.21	−.12	−.24	−.06
9	−.09	−.01	.05	.03	−.13	.13	.14	−.23		.13	.02	−.07	.07	.04	.05
10	.10	.30	.14	.21	.24	.20	.20	−.21	.13		.07	.28	.15	.29	.06
11	.29	.25	.06	.28	.02	−.05	−.01	−.22	.02	.07		.23	.09	.02	−.06
12	.19	.22	.21	.51	.14	.14	.21	−.21	−.07	.28	.23		.15	.12	−.01
13	.13	.17	−.18	.20	08	.07	.22	−.12	.07	.15	.09	.15		−.10	.12
14	.08	.09	.28	.09	.08	.17	−.09	−.24	.04	.29	.02	.12	−.10		.04
15	−.11	−.13	−.25	.07	−.01	−.08	.11	−.06	.05	.06	−.06	−.01	.12	.04	

virtually identical for the two solutions and led to the same factor interpretations.

Factor A (*Educational Background and Achievement*). The three major positive loadings were on the primaries School Achievement, Higher Educational Achievement, and Educational-Vocational Consistency. There was a lesser negative loading on Stability. The factor was defined as: "Educational achievement in childhood and adult life and its application to real-life situations."

Factor B (*Upward Mobility and Drive*). The three highest loadings were a positive on the primary Drive, and negatives on Vocational Satisfaction and Parental Family Adjustment. The combination of a positive loading on the Drive factor and a negative loading on the Vocational Satisfaction factor was to be expected in terms of their respective definitions. It has been pointed out that the Vocational Satisfaction factor hints of satisfaction achieved rather as a result of limited goals than of special application or drive. The negative association between the Drive and Parental Family Adjustment factors was predicted by one of the psychologists associated with the study. McMurry[4] claims that

[4] R. N. McMurry, personal communication, 1965.

TABLE 2

ORTHOGONAL AND OBLIQUE ROTATED MATRICES FOR THE 5 SECOND-ORDER FACTORS

Primary factor	Orthogonal matrix					Oblique matrix				
	Second-order factors					Second-order factors				
	A	B	C	D	E	A	B	C	D	E
1	*.73*	−.14	.15	.11	.06	*.67*	−.11	.00	.01	.05
2	*.57*	.04	.10	.25	−.06	*.53*	.06	−.05	.23	−.06
3	.09	*−.52*	*.32*	*.35*	.00	−.01	*−.51*	.28	.28	.04
4	.19	.07	*.69*	.00	−.17	.06	.04	*.60*	−.09	−.06
5	*.30*	.00	.00	*.61*	.19	.26	.05	−.10	*.60*	.20
6	−.03	−.05	.05	*.51*	−.08	−.06	−.04	−.01	*.50*	−.05
7	−.07	*.41*	.21	.07	.02	−.10	*.40*	.19	.06	.07
8	*−.32*	−.03	−.21	.17	*.68*	*−.30*	.01	−.06	.21	*.66*
9	−.08	.10	−.01	.04	*−.34*	−.06	.08	−.04	.04	*−.32*
10	.12	.19	.21	*.46*	*−.33*	.08	.18	.08	*.43*	−.27
11	*.38*	−.01	.26	−.11	−.07	*.33*	−.01	.17	−.15	−.05
12	.14	.10	*.69*	.18	−.03	.00	.09	*.61*	.09	.09
13	.21	*.49*	.12	.07	−.03	.20	*.49*	.04	.07	.00
14	.01	−.26	.10	.28	*−.40*	−.02	−.27	.03	.25	*−.38*
15	−.11	*.30*	−.01	−.05	−.12	−.09	.29	−.01	−.03	−.11

Note.—Loadings of .30 or greater are italicized.

over the years of his consulting experience in industrial organizations, he has found that happy and adjusted parental family situations often produce pleasant or easy-going people but do not appear to generate the personal striving and aggressiveness necessary for upward mobility in the line management hierarchy. The factor was defined as: "Inner drive to be outstanding in performance, to set high goals, and to move upward in the organizational hierarchy."

Factor C (Personal-Social Leadership). The three highest loadings were on the primaries Leadership and Group Participation, Vocational Decisiveness, and Drive. The common element indicated by these loadings appeared to be the action-oriented decisiveness and personal leadership in the social and occupational environment. The factor was defined as: "Participation and personal leadership in social situations and occupational environments."

Factor D (Financial Achievement and Background). The two major loadings were on the primaries Financial Responsibility and Early Family Responsibility with lesser loadings on Professional-Successful Parents and Drive. The factor was defined as: "High achievement and responsible management of finances and early marriage with sole responsibility for family support, associated with a parental family background of affluence and success."

Factor E (Stability and Status Quo Orientation). The major loading was on the Situational Stability factor which indicated established security and concern over the maintenance of what had been achieved. The rather restrictive flavor of this factor was emphasized by the negative loadings on Selling Experience, School Activities, and Professional-Successful Parents, that is, primaries which indicated mobility and action. The factor was defined as: "Establishment and maintenance of stability in the occupational and home environments associated with avoidance of outside activities."

A clear structure was obtained for the orthogonal rotation, and, as may be expected under these circumstances, the correlations between the oblique second-order factors were quite low. These correlations are given in Table 3.

TABLE 3

Correlations between the Oblique Second-Order Factor Axes

	A	B	C	D	E
A		−.06	.39	.13	−.01
B	−.06		.06	−.07	−.12
C	.39	.06		.27	−.29
D	.13	−.07	.27		−.10
E	−.01	−.12	−.29	−.10	

Analysis of Variance

An analysis of variance was undertaken using both the derived beta-weight and the unit-weight scores for the 15 oblique first-order factors across the 10 occupational groups. Examination of Table 4 will show that virtually all factors discriminated between the occupational groups at the .001 level of confidence or better, though some of the factors were clearly more efficient in this respect than others. The discriminative power of the factors was further investigated by a complete series of *t* tests of significance of the differences between the mean unit-weight factor scores for each of the 45 possible comparisons of pairs of occupational groups.[5]

Each factor showed some differences for pairs of occupational groups significant at the .05 level or better. These differences range from 8 for the Parental Family Adjustment factor to 37 out of a maximum of 45 for the Financial Responsibility factor, with a mean of 25.8 per factor. Each pair of occupational groups showed some differences at the .05 level or better ranging from 3 for the comparison between district sales managers and other line executives to 14 out of a maximum of 15 between sales and hourly personnel with a mean of 8.6 significant differences per occupational comparison.

Identification of Factors with Most Potential for Operational Use

Although the beta- and unit-weight scores for the 15 first-order factors will be routinely

[5] This table of results is available from the Industrial Relations Center, University of Chicago, upon request.

TABLE 4

ANALYSIS OF VARIANCE FOR THE UNIT- AND BETA-WEIGHT SCORES FOR THE 15 OBLIQUE FIRST-ORDER FACTORS ACROSS 10 OCCUPATIONAL GROUPS

Factor	Beta-weight scores		Unit-weight scores	
	F	*r*	*F*	*r*
1	15.71**	.27	13.25**	.39
2	18.08**	.44	45.08**	.61
3	4.16**	.23	6.26**	.28
4	12.53**	.38	26.30**	.51
5	18.42**	.45	43.25**	.61
6	5.28**	.26	28.37**	.53
7	1.41	.14	2.80*	.19
8	16.52**	.43	157.39**	.82
9	3.24*	.20	14.55**	.40
10	5.47**	.26	13.68**	.39
11	8.08**	.31	20.40**	.46
12	4.64**	.24	4.60**	.24
13	3.85**	.22	29.78**	.53
14	28.82**	.53	27.88**	.52
15	3.07*	.20	5.24**	.26

Note.—$df = 9/670$.
* $p < .01$.
** $p < .001$.

calculated for other research purposes for some years to come, attention will also be directed toward the identification of those factors which seem to offer the most potential for operational use in terms of their reliability and construct validity. These latter factors will be subject to specific research studies in organizational settings.

In a practical situation, the use of beta-weight scores is extremely cumbersome if not unfeasible when computer facilities are not available. The use of unit-weight scores immediately raises questions concerning the number of items assigned to each factor, the correlations between the respective unit-weight and beta-weight factor scores, and the intercorrelations for the unit-weight scores. This information is given in Table 5.

The internal-consistency reliability estimates obtained by applying the Kuder-Richardson formulas to the unit-weight factor scores are given in Table 6. The figures point up the deficiencies of factors such as 12, which is defined by a small number of items, and of 13 and 14, where the factorial structure is unclear. Although these latter factors have a sizable number of items allocated to them, only three or four items per factor have loadings of .30 or larger with the remaining loadings being indeterminate.

The ability of the unit-weight factor scores to discriminate between occupational groups was discussed in connection with Table 4 and *t*-test analysis. In this connection it is of interest to note that although Factor 15 (General Health) has a clear factorial structure with at least a moderate number of items with significant loadings, it has relatively low reliability, and is relatively weak in its power to discriminate between occupational groups. The final selection of factors must await the results of studies relating the factor scores to successful performance within an occupational group. On the basis of presently available evidence, however, it

TABLE 5

PRODUCT-MOMENT INTERCORRELATIONS OF THE UNIT-WEIGHT SCORES FOR THE 15 FIRST-ORDER OBLIQUE FACTORS AND CORRELATIONS BETWEEN THE RESPECTIVE BETA- AND UNIT-WEIGHT FACTOR SCORES (β-u)

Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		−.10	−.04	−.07	−.11	.04	.08	.02	.49	.09	.19	.10	.15	−.04	.02
2	−.10		−.01	.06	−.05	.05	.12	.05	.19	.22	.25	.01	.25	.03	−.10
3	−.04	−.01		.12	.08	.07	−.07	−.07	.07	−.03	−.07	−.12	−.01	.53	−.03
4	−.07	.06	.12		.05	.01	.13	.01	.54	.19	.16	.03	.26	.08	.20
5	−.11	−.05	.08	.05		.35	−.01	−.14	.04	.07	.07	.06	−.14	.01	−.03
6	.04	.05	.07	.01	.35		.01	.04	.08	.05	06	−.03	−.06	.21	−.06
7	.08	.12	−.07	.13	−.01	.01		−.07	−.06	−.11	−.02	.25	.24	.02	.06
8	.02	.05	−.07	.01	−.14	.04	−.07		−.19	−.13	−.06	.07	−.42	.14	−.12
9	.49	.19	.07	.54	.04	.08	−.06	−.19		.34	.12	−.02	.23	.10	.18
10	.09	.22	−.03	.19	.07	.05	−.11	−.13	.34		.02	.15	.22	.10	.05
11	.19	.25	−.07	.16	.07	.06	−.02	−.06	.12	.02		.17	.34	.03	−.01
12	.10	.01	−.12	.03	.06	−.03	.25	.07	−.02	.15	.17		.01	−.11	.01
13	.15	.25	−.01	.26	−.14	−.06	.24	−.42	.23	.22	.34	.01		.16	.07
14	−.04	.03	.53	.08	.01	.21	.02	.14	.10	.10	.03	−.11	.16		.02
15	.02	−.10	−.03	.20	−.03	−.06	.06	−.12	.18	.05	−.01	.01	.07	02	
β-u	.87	.83	.86	.77	.83	.88	.90	.76	.72	.80	.86	.87	.72	.64	.87

would seem that the first eight factors presented in the interpretations offer the most potential for operational use.

DISCUSSION

The vast bulk of previous research investigating the validity of quantifiable items of background data in predicting occupational success has taken the approach of determining the extent to which individual items could differentiate between successful and unsuccessful performance within a particular occupational group and the subsequent construction of an index based on weighted or unweighted responses to the discriminating items. This research has rightly been criticized on the grounds of empiricism, its failure to provide any explanation as to why certain background items should be combined or be predictive, and because of its failure to shed any light on the dynamic relationships between early home and school environments and eventual performance in an occupation.

The present research followed the approach of first attempting to identify and define some of the significant underlying dimensions of background data. This was implemented through successive factor analyses of responses of an occupationally heterogeneous sample of 680 male employees to a wide spectrum of background-data items derived from the home, school, college, and work environments. The final factoring employed the principal axis solution and yielded 15 factors which accounted for 66.6% of the correlation and 43.3% of the variance. The factors were rotated both to orthogonal simple structure using the equamax criterion (Saunders, 1962) and to oblique simple structure using the promax criterion (Hendriksen & White, 1964). All 15 factors from both solutions were interpretable but the structure was clearer for the oblique solution.

A second-order analysis of the correlations between the oblique primary-factor axes yielded five factors with very similar factorial structures for both the orthogonal and oblique rotational solutions. This would seem to indicate that, in the second order, the background-data factors were largely uncorrelated. The combinations of primaries represented in the second-order factor were logically and theoretically acceptable.

An analysis of variance utilizing both unit-weight and beta-weight scores for the oblique primary factors across the 10 occupational groups comprising the sample indicated that virtually all factors would discriminate between the occupational groups at the .05 level of confidence and the majority of them at the .001 level or better. The discriminatory ability of the factors was investigated through the calculation of the significance of the differences between the mean unit-weight factor scores for each of the 45 possible pairs of occupational groups. The results indicated that each pair of occupational groups showed some differences significant at the .05 level of confidence or better and, conversely, that each factor showed some significant differences for pairs of occupational groups.

A tentative selection was made of eight factors which seemed to have the most potential for operational use on the basis of clearness of factor definitions, the number of items defining the factor, the correlation between respective unit-weight and beta-weight scores, and the Kuder-Richardson reliability coefficients. Attention should also be directed toward the second-order factors which had the advantage of being uncorrelated

TABLE 6

KUDER-RICHARDSON RELIABILITY COEFFICIENTS FOR THE UNIT-WEIGHT FACTOR SCORES

Factor number	*M*	No. items	*SD*	Reliability		
				K-R 8	K-R 20	K-R 21
1	2.90	6	1.64	.71	.57	.53
2	5.34	11	2.40	.68	.60	.58
3	4.42	8	2.11	.75	.65	.64
4	3.43	8	2.07	.73	.64	.62
5	4.43	8	2.22	.77	.69	.68
6	7.78	13	2.80	.77	.72	.65
7	6.83	10	2.18	.73	.65	.60
8	5.16	9	2.47	.82	.76	.72
9	3.42	7	2.07	.75	.64	.55
10	3.27	7	1.78	.68	.55	.52
11	2.90	4	1.64	.87	.54	.52
12	2.21	5	1.36	.66	.43	.41
13	5.91	12	2.24	.62	.51	.44
14	5.41	10	2.01	.60	.47	.43
15	2.76	5	1.31	.66	.43	.34

and of seeming to represent five broad behavior patterns identified as:

A. Educational Background and Achievement
B. Upward Mobility and Drive
C. Personal-Social Leadership
D. Financial Achievement and Background
E. Stability and Status Quo Orientation

The research has provided a framework for the interpretation and examination of some significant behavioral dimensions inherent in background-data items for the occupations represented in the sample. It has shown that these dimensions as represented by the primary factors will differentiate between *S*s in different occupational groups. Since the correlation between background-data scores and the skill and ability measures which traditionally discriminate between occupational groups has generally been found to be low, the tentative conclusion may be drawn that the background-data scores are intrinsically discriminative for occupational classification.

The extent to which the background-data factors will be predictive of successful performance within an occupational group can be investigated by multiple-regression analyses against selected empirical criteria. Such investigations may well lead to the identification of characteristic background profiles for successful performers in different occupational groups.

REFERENCES

BAEHR, M. E., BURNS, R. K., & MCMURRY, R. N. *Personal history index.* (Research ed.) Chicago: Industrial Relations Center, University of Chicago, 1963.

BAEHR, M. E., BURNS, R. K., & MCMURRY, R. N. *Personal history index.* (Rev. ed.) Chicago: Industrial Relations Center, University of Chicago, 1965.

BUEL, W. D. Biographical data and the identification of creative research personnel. *Journal of Applied Psychology,* 1965, **49**, 318–321.

DUNNETTE, M. D., & MAETZOLD, J. Use of the weighted application blank in hiring seasonal employees. *Journal of Applied Psychology,* 1955, **39**, 308–310.

DWYER, P. S. The determination of the factor loadings of a given test from the known factor loadings of other tests. *Psychometrika,* 1937, **2**, 173–178.

ELLIOTT, L. L. Prediction of success in WAF basic training by two background inventories. *USAF WADD Tech. Note,* 1960, No. 60–216.

GLENNON, J. R., ALBRIGHT, L. E., & OWENS, W. A. *A catalog of life history items.* Chicago: American Psychological Association, Division 14, Scientific Affairs Committee, 1963.

GOLDSMITH, D. B. The use of the personal history blank as a salesmanship test. *Journal of Applied Psychology,* 1922, **6**, 149–155.

GUTHRIE, M. R. The measurement of personal factors related to success in office workers. *Journal of Applied Psychology,* 1956, **40**, 87–90.

HENDRIKSEN, A. E., & WHITE, P. O. Promax: A quick method of rotation to oblique structure. *British Journal of Statistical Psychology,* 1964, **17**, 65–70.

KERR, W. A., BAEHR, M. E., & BURNS, R. K. *The biographical index.* Chicago: Industrial Relations Center, University of Chicago, 1961.

LEVINE, A. S., & ZACHERT, V. Use of biographical inventory in the Air Force classification program. *Journal of Applied Psychology,* 1951, **35**, 241–244.

LOCKWOOD, H. C., & PARSONS, S. O. Relationship of personal history information to the performance of production supervisors. *Engineering Industrial Psychology,* 1960, **2**, 20–26.

MCMURRY, R. N. Validating the patterned interview. *Personnel,* 1947, **23**, 263–272.

MINOR, F. J. The prediction of turnover of clerical employees. *Personnel Psychology,* 1958, **11**, 393–402.

MORRISON, R. F., OWENS, W. A., GLENNON, J. R., & ALBRIGHT, L. E. Factored life history antecedents of industrial research performance. *Journal of Applied Psychology,* 1962, **46**, 281–284.

MOSEL, J. N., & WADE, R. R. A weighted application blank for reduction of turnover in department store sales clerks. *Personnel Psychology,* 1951, **4**, 177–184.

SAUNDERS, D. R. Trans-varimax: Some properties of the ratiomax and equamax criteria for blind orthogonal rotation. *American Psychologist,* 1962, **17**, 395–396. (Abstract)

SCOLLAY, R. W. Validation of personality history items against a salary increase criterion. *Personnel Psychology,* 1956, **9**, 325–336.

SEAQUIST, M. R., BARRY, J. R., & SELLS, S. B. Adaptability screening of flying personnel; life history inquiry approach based on the personal history and background information questionnaire. *USAF School of Aviation Medical Report,* 1956, No. 56-45.

SMITH, W. J., ALBRIGHT, L. E., & GLENNON, J. R. The prediction of research competence and creativity from personal history. *Journal of Applied Psychology,* 1961, **45**, 59–62.

SUPER, D. E. The biographical inventory as a method for describing adjustment and predicting success. *Bulletin de l'Association Internationale de Psychologie Appliquée,* 1960, **9**, 18–39.

(Received December 1, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 491–496

RESPONSE TENDENCIES IN THE SVIB:

THE POPULAR, THE RARE, AND THE SOCIALLY DESIRABLE

DONALD G. ZYTOWSKI AND JAMES A. WALSH

Iowa State University

A number of response sets: popular, rare, socially desirable, liking, and indifference were examined for their correlations with each other and personality social desirability (SD), and for their influence on the scales of the SVIB-M. It was found that the scale of SD formed from personality items is independent of socially desirable interest scales, and that the latter vary more freely from item probability of endorsement. Certain of the scales for which the SVIB is scored vary fairly directly with 1 or more response tendencies. The significance of these findings for interpretation of the profile sheet is discussed.

Inventories of affective responses (interest, attitude, and personality) appear to be susceptible in varying amounts to a variety of response tendencies. A number have been identified: social desirability (SD), tendencies to avoid or endorse extreme statements, acquiescence, and the like. Some scales of response tendencies have been constructed to aid in the interpretation of other descriptive scales, as in the case of the *L*, *F*, and *K* scales of the MMPI. Their effectiveness in achieving this purpose is questionable.

Of the three kinds of inventories, interest inventories seem to have escaped close examination for the effects of response bias. Only the Kuder inventories are scored for the V (popular response) scale, which could be said to be a measure of response tendency, though it is intended as an index to the validity of the interest-area scales. In the case of the Strong Vocational Interest Blank (SVIB), Filbeck and Callis (1961) have constructed a V scale (using rare responses) after the fashion of Kuder. Although they correlated the V scale with the occupational scales of the SVIB, it was not conceived as an opportunity to assess the contribution of a tendency to elect unpopular responses to the occupational scales. Berdie (1943) has investigated the effects of responding exclusively in terms of like (L), indifferent (I), or dislike (D) to the SVIB, by showing score profiles generated from answer sheets marked with one response only. It is not clear how he marked the Items 281–400, which do not use the LID format, and he did not analyze the relative contributions of each of these sets to the variance in the occupational scales.

Using a different approach to the problem, Walsh and Keasey (1966) investigated the relationship of a number of personality- and attitude-scale response tendencies to scores on the SVIB. In general, their findings showed factors composed mainly of occupational interests as distinct from factors emerging from response tendencies. That is, identifiable response tendencies did not tend to "get through" to the scores on the occupational scales. They did find that Edwards' (1957) Social Desirability (*SD*) scale composed of items from personality inventories (*PSD*) had moderate positive correlations (median $r = .27$) with SVIB Group V scales. The tendency to describe one's self in terms of desirable personality attributes is positively, but very moderately, related to similarity with social service occupations among men. The relationship did not hold for women.

The contrast of Walsh and Keasey's finding of response scales from items external to interest inventories as essentially unrelated to interest scores, with Berdie's (1943) and Filbeck and Callis' (1961) findings that response sets formed from interest items do bear considerable influence in the occupational scales, raises some intriguing questions concerning the generality of response tendencies from one affective domain to another. Fur-

ther evidence of this kind exists in Taylor's (1961) finding of a .58 correlation between *PSD* scores and an SD scale formed exclusively of items from attitude scales. Thus the focus of the present investigation was to assemble an SD scale composed of SVIB items, and to compare that scale with Edwards' *PSD* scale and with other response tendencies unique to the SVIB as measured by scales composed of its own items.

Construction of the Interest Social Desirability Scale

The construction of an SD scale from interest items was undertaken in the method normally employed by Edwards (1957). Eighty-nine volunteers from an undergraduate introductory class in psychology rated the social desirability of endorsing (responding "like") all items of the 1966 revision of the SVIB (which includes all the items from the previous version now used in the occupational scales). A 9-point scale was employed, ranging between "extremely socially desirable" and "extremely socially undesirable" with "neither desirable nor undesirable" at the center.

The items from 281 on, which do not receive LID responses, were rated for the social desirability of endorsing them, much as MMPI items are rated. In Part VII, which calls for the preference between two items, both response alternatives were rated. Means and standard deviations of the 9-point ratings were computed for each item. The mean rating was taken as the index of SD, and the standard deviation was taken as reflecting consensus.

Typical of items for which endorsement was regarded as very socially desirable were

TABLE 1

CORRELATIONS BETWEEN ITEM SDs AND p OF L, I, OR D RESPONSE

	σ *SD*	pL	pI	pD	$\bar{X}$	σ
SD	−.35	.43	−.19	−.38	6.09	1.01
σ SD		−.02	−.04	−.05	1.41	.22
pL			−.47	−.86	.371	.17
pI				−.03	.351	.09
pD					.278	.15

Note.—N = 288 items.

No. 93 (Surgeon), No. 69 (Physician), and No. 311 (President of a society or club). Socially *un*desirable was endorsement of such items as No. 292 (steadiness and permanence of work most important), No. 393 (making excuses when caught in a mistake), and No. 13 (liking to be an auto mechanic). Consensus was strong on items with extreme ratings, as would be expected. Items on which agreement concerning social desirability was poor were represented by No. 221 (expressing opinions openly) and No. 107 (Calculus).

Two opportunities were available to check the validity of this procedure of rating for SD. One was to compare the SD indexes of occupations in Part I with those rated in the North-Hatt (1953) study of occupational prestige. For 64 occupations, the correlation between the two forms of ratings was .85. Campbell [1] supplied the authors with SD ratings on a 5-point scale of the first 100 items of the SVIB (occupations) given by undergraduate psychology students at the University of Minnesota. The correlation between these ratings and the present ones is .91. It is assumed that the SD ratings of SVIB items obtained for this study are valid and representative of those which might be given by other populations of similar age and education.

Correlations between item SD ratings and probabilities of endorsement were computed and are given in Table 1.

The correlation between index of SD and probability of a "like" response for the 288 items common to Forms M and T-399 is .43. Though a relationship of this magnitude has only the slightest possibility of occurring by chance, it does not approach that (.87) found by Edwards (1957) between SD and probability of endorsement of personality items. This suggests that, on the whole, a person is relatively free to report what objects and activities he likes and dislikes, the group norms of desirability of objects and activities which may exist notwithstanding. Reports of interests are less influenced by SD than endorsements of personality descriptions. One might hypothesize that what is desirable is more independent of what is popular in

[1] D. P. Campbell, personal communication, 1966.

interest inventories than in personality inventories.

Of additional interest is the correlation between L and D responses. It appears that if L is not employed, the response is much more likely to be D than I. This large negative correlation between L and D also accounts for the similarity between the SD–L correlation and the negative SD-D one.

Although the independence of socially desirable responses and most likely responses suggests the possibility of construction of a Strong type—contrasting groups—scale of Interest SD, a simple scale of the 20 most- and 20 least-desirable items was assembled to provide comparisons. Because of the substantial negative correlation between L and D responses, it was assumed appropriate to score 1 point for liking the high SD items and the same for disliking the low SD items, rather than to use differential positive and negative weights. This scale was identified as Interest Social Desirability No. 1 (ISD No. 1).

It was further decided, for heuristic purposes, to form another ISD scale using the method of contrasting groups, as is typical of Strong's scales for his blank. For this purpose, it was necessary to employ the item SD in a manner which would obtain the *socially desirable response* for each item. A Strong-type scale was assembled as follows: items were arranged serially by magnitude of SD rating. The series was then divided in the same proportions as are given L, I, and D responses over all items by men in general (MIG). Thus for an item whose SD index was in the top 38% of the series, the socially desirable response was taken to be L, while if it fell in the middle 35% the socially desirable response was I, and in the bottom 28% of the SD series it was assigned D as the socially desirable response. For every item, the response which received the greatest proportion of endorsement by the MIG group was labeled as the *most likely response*. A scale was formed by assembling all the items whose most desirable and most likely responses differed by two categories, making a total of 33. ISD No. 2 was scored as the number of endorsements of these more desirable than likely responses.

TABLE 2

INTERCORRELATION MATRIX OF 6 RESPONSE TENDENCIES ON THE SVIB-M

	ISD No. 1	ISD No. 2	V	L	I	$\bar{X}$	σ
PSD	−.07	−.12	−.26	.16	.28	29.44	6.7
ISD No. 1		.51	.08	−.03	−.21	16.47	4.5
ISD No. 2			.29	−.12	−.16	8.00	3.5
V				−.09	−.26	9.8	5.0
L					−.51	98.95	32.5
I						103.63	34.7

With two forms of Interest SD scales a sembled, and with Filbeck and Callis' (1961 V scale standing for the tendency to endor rare responses, it was possible to score number of SVIB answer sheets and correla the response-set scores with those normall provided on the SVIB profile. Walsh ar Keasey (1966) made the Form M answ sheets from their study available for this pu pose, providing Edwards' *PSD* scores ar L and I response frequencies. The answ sheets were handscored for the new respon scales, and the data punched on cards f analysis on an IBM 360–50 computer. C first interest are the intercorrelations betwe the response tendencies themselves. They a presented in Table 2. D was excluded becau it is virtually the mirror image of L.

Several observations can be made. Mo notable is the correlation indicating that t tendency to describe one's self in socially d sirable personality terms (*PSD*) is essential unrelated to the tendency to endorse sociall desirable interests (ISD), no matter ho the latter are defined. However, SD tendenci are somewhat more strongly negatively r lated to the tendency to endorse rare item and positively to the tendency to emplo "indifferent" responses. It might also be note that the significant negative correlation b tween I responses and the V (rare respons scale is a spurious function of the lack of items on the V scale. That is, if a pers endorses many I items, he will be less ab to choose items which contribute to the scale.

Turning to the interrelationships of t two ISD scales, it can be seen that the E wards-style scale of most- and least-desiral items is significantly related to the Stro style scale, but the correlation is low enou

TABLE 3

Correlations of Several Measures of Response Tendency with 49 SVIB-M Scales

Scale	*PSD*	ISD No. 1	ISD No. 2	V	L	I
Group I						
Artist		−22*		28**		−21*
Psychologist						23*
Architect		−25**	−20*			
Physician		−20*	−24*			
Osteopath					33**	
Dentist			−21*			
Veterinarian		−34**	−23*	−39**	33**	
Group II						
Mathematician		−24*	−24*			
Physicist		−26**	−36**			
Engineer			−42**			
Chemist			−35**			
Group III						
Production Manager			−25**			
Group IV						
Farmer		−55**	−52**	−26**		
Aviator		−38**	−49**	−28**		
Carpenter		−47**	−52**	−26**	22*	
Printer		−43**	−48**	−27**	25**	
Math–Science Teacher	20*	−25**	−44**	−44**	33**	32**
Industrial Arts Teacher		−45**	−57**	−38**	25**	24*
Vocational Agricultural Teacher		−53**	−55**	−49**		28**
Policeman	20*	−21*	−30**	−47**	30**	34**
Forester	20*	−46**	−57**	−46**	23*	25**
Group V						
YMCA Physical Director	28**			−46**	32**	42**
Personnel Director	22*	20*				45**
Public Administrator	25**			−26**		49**
YMCA Secretary	26**			−26**	26**	47**
Social Science HS Teacher				−29**		37**
City School Superintendent	32**					50**
Social Worker	21*					41**
Minister	20*					35**
Group VI						
Musician		−24*				
Group VII						
CPA	−25**	45**	21*		−25**	−21*
Group VIII						
Senior CPA			−29**	−40**		
Accountant		23*				
Office Man				−21*	31**	20*
Purchasing Agent		25**				
Banker						
Mortician			24*			
Pharmacist		21*	25**		22*	
Group IX						
Sales Manager		34**	29**			
Real Estate Salesman			38**			
Life Insurance Salesman		34**	47**			
Group X						
Advertiser			40**	−34**	−22*	
Lawyer			45**	28**	−27*	
Author–Journalist			27**	40**	−23*	
Group XI						
President Manufacturing Concern		31**	21*	32**		−28*
Specialization Level	32**					36**
Interest Maturity	50**					45**
Occupational Level	26**	24*	28**	33**	37**	
Masculinity–Femininity	25**	−28**	−49**	−20*		

* $p < .05$.
** $p < .01$.

to preclude the assumption that they are measuring the same variable.

Table 3 shows the correlations between the six response tendencies and all scales of the SVIB-M. In interpreting this table, it should be remembered that occupations within groups are more highly related to each other than to occupations outside a given group. Thus, if a response tendency correlates to an extent with one scale in a group, it is likely to correlate similarly with all in a group. Correlations with a $p > .05$ have been omitted.

First glance at Table 3 should suggest that the Group IV scales of technical occupations seem to be exceptionally vulnerable to response sets, especially to endorsement of interest-measure defined socially undesirable and rare responses. The several Teacher scales in this group and Policeman and Forester scores are also noticeably affected by the tendencies to endorse L and I responses. Since these correlations appear on two of the three response alternatives, one may say that scores on these scales covary with the tendency to avoid responding in terms of dislikes. None of the other SVIB groups appears to be so highly correlated with so many kinds of response tendencies. This finding is in accord with those of Walsh and Keasey (1966).

Taking each scale of response bias in order, some observations are appropriate as to its main area of influence. As already noted, the tendency to describe one's personality in SD terms has a moderate relationship with social service similarities. But also notable is this scale's relationship ($p < .01$) with all of the nonoccupational scales. Increasing interest maturity, particularly, is associated with socially desirable self-description (*PSD*) but not with endorsement of SD interests. The same is true of the Occupational Level and Masculinity-Femininity scales.

The tendency to endorse socially desirable interests, especially those more desirable than popular, tends to make one look like persons in persuasive and verbal occupations and unlike science and technology occupations. This antithetical relationship between these groups of interest similarity is visible in other reports, for instance, Darley and Hagenah's (1955) frequency tables of profile patterning. The Edwards-type ISD scale has correlations with the occupational scales similar to the Strong-type ISD measure, except in Group X. There, the positive correlations on the V scale tend to repeat the stronger correlations of the desirable-but-rare scale.

Here it might be asked if the component of SD interests in any occupation is related to the degree of prestige which is accorded the occupation itself. The correlation between the North-Hatt (1953) index of occupational prestige and the correlations of ISD No. 1 with the SVIB occupations is .13—not significant for an n of 34 occupations common to both scales. This suggests that the prestige of an occupation is not dependent upon the prestige of the interests associated with it, or that socially desirable interests are not sufficient if endorsed to produce similarity with prestige occupations.

The tendency to select rare responses shows up as influencing scores in verbal occupations, as noted above, and also in the artist and president of a manufacturing concern scales. The converse applies to Group IV and some social service scales: high scores on the rare set scale tend to be associated with low scores in these occupational scales.

While Berdie (1943) shows that answer sheets filled out with all L responses enhance social service (Group V) scores, the present data show that the technical occupations (Group IV) covary to a greater extent with actual use of L responses than does Group V. The latter group covaries quite directly with I responses, as do the Teacher scales within the technical group. In addition, the I tendency forms a small but significant portion of variance in the Specialization Level and Interest Maturity scales. The OL and MF scales also show correlations with all of the SD scales and the V scale, but MF is inversely related to all sets but PSD.

The one occupational group which seems least affected by any response tendency studied here is VIII, Business Detail. There is no interpretable pattern of correlations within this group, and the largest proportion of correlations are just over the .05 level of probability. Those which do appear may be the consequence of random variation.

Conclusions and Implications

Several conclusions may be drawn from the findings of this study. For one, response tendencies unique to interest inventories (L, I, and D) can be identified and measured. Second, tendencies originally identified in other affective domains (socially desirable and rare or popular) can be constructed within the interest domain, and in the case of SD are not equivalent with the same tendency in the personality domain. Further, the correlation between item SD and probability of endorsement in order of descending magnitude is: personality, attitude, and interest. Still the correlation on the SVIB is considerably larger than could be expected to appear by chance.

While Walsh and Keasey (1966) did not find any *major* associations of variance from response sets associated with personality and attitude scales with SVIB scales, the findings of the present study appear to support the conclusion that response sets formed within interest items do "get through" in varying amounts to many of the occupational scales. In many, the correlations are larger than chance, but minimal in terms of interpretation. On the other hand, 20–25% of the variance in Group V scores can be accounted for in terms of the tendency to endorse the I response. Slightly less variance in the Group II and IV scales is captured by a negative relationship to the use of socially desirable responses, and in Group X a large portion of a scale score may be the consequence of choosing socially desirable but rare interests while avoiding the liking response.

An important practical question is whether or not the tendency to respond in accordance with a consistent set or bias makes any difference in the occupational scale scores. After all, the SVIB is an empirical set of scales, normed on a variety of occupations, and certain tendencies to respond may be a part of the essence of the occupation captured in any scale. Granting this, but given a person who has, say, all of the interests appropriate to a Group IV occupation, but who also tends to endorse socially desirable interests, enough may be taken off his occupational score to reduce it to an equivocal level. In the light of this prospect, it is important to understand just how certain response sets contribute to scores on the Strong Blank.

REFERENCES

BERDIE, R. F. Likes, dislikes, and vocational interests. *Journal of Applied Psychology*, 1943, **27**, 180–189.

DARLEY, J. B., & HAGENAH, T. *Vocational interest measurement.* Minneapolis: University of Minnesota Press, 1955.

EDWARDS, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957.

FILBECK, R. W., & CALLIS, R. A verification scale for the SVIB, men's form. *Journal of Applied Psychology*, 1961, **45**, 318–324.

NORTH, C. C., & HATT, P. K. Jobs and occupations: A popular evaluation. In R. Bendix & S. M. Lipset (Eds.), *Class, status, and power.* New York: Free Press, 1953. Pp. 411–426.

TAYLOR, J. B. What do attitude scales measure: The problem of social desirability. *Journal of Abnormal and Social Psychology*, 1961, **62**, 386–390.

WALSH, J. A., & KEASEY, C. Stylistic and response set variance in the SVIB: Fact or fiction? In D. P. Campbell (Chm.), Strong Memorial Symposium on interest measurement. Symposium presented at the University of Minnesota, May 1966.

(Received December 5, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 497–502

DIFFERENCES IN OPINION-SURVEY RESPONSE PATTERNS AS A FUNCTION OF DIFFERENT METHODS OF SURVEY ADMINISTRATION

JOHN R. HINRICHS[1]
IBM Corporation
White Plains, New York

AND

ROBERT D. GATEWOOD
Purdue University

For a population of male technical employees in a large national organization the method of responding to an opinion survey made no difference in the level of responses given. However, the conditions under which the survey was administered did have an effect on response. When employees were surveyed on their job locations under the supervision of a company representative there was a tendency to respond more favorably to a significant number of general-opinion questions, particularly questions dealing with the "company in general," than when they were permitted to respond in a nonoffice location. The underlying dynamics of this result are not clear, but the significant point is that the method of questionnaire administration can serve as a potential source of distortion in employee-opinion-survey data.

Today the employee opinion survey is a widely used tool available to managers of large industrial organizations for evaluating significant aspects of their company's operations and to social scientists studying the interaction of individuals and organizations. The use of opinion surveys in industry has increased markedly in recent years, and in many companies data from these types of surveys find their way into a variety of management decision-making considerations. Surveys are utilized to audit management practices, to evaluate the state of employee morale, to understand the dynamics of work-group motivation, to evaluate training needs, to audit the effectiveness of differing supervisory and leadership strategies, to clarify channels of communications, and so forth. Through many years of experience and continued refinement and polishing, the opinion survey has been developed into a reliable and valid technique for evaluating a spectrum of organizational problems.

However, in contrast to the great deal of attention paid to refining and polishing the content of opinion surveys (i.e., the actual questions which are asked), there has been relatively little attention to the considerations surrounding the mechanical administration of surveys. A few studies have evaluated differences in attitude-survey responses collected under conditions of complete respondent anonymity versus conditions in which respondents were identified on their questionnaires. In reviewing this literature, Rosen (1960) concludes that, while statistically significant differences have been found in a number of these studies, "most of the researchers state that the differences between anonymous and identified subjects were of little practical importance [p. 676]." The 10 studies which he cites cover groups in academic, military, or community settings. Rosen concludes that there is not apt to be serious distortion as a result of respondent identification in situations where the survey is conducted under less than highly threatening circumstances.

Pelz (1959) arrives at a similar conclusion following his study in a government agency. He found practically no differences in survey data collected under conditions of full anonymity versus identification with assurances that replies would be confidential among the researchers who came from an outside survey organization. Pelz termed both of these conditions as relatively unthreatening.

Somewhat in contrast to the rather mild effects noted in these nonindustrial studies, however, two studies in industrial settings suggest that significant distortion can occur under different survey-administration conditions, and that this distortion increases with

[1] Now at Cornell University, Ithaca, New York.

the degree of perceived threat to the individual. Klein, Maher, and Dunnington (1967) found that identified survey respondents in a manufacturing environment registered more favorable attitudes than anonymous respondents, particularly on items dealing with salary and with top management. These differences were more marked under conditions of high threat (*S*'s immediate manager designated whether he would be in the identified or anonymous subsample) than was the case under conditions of low threat (*S*s assigned randomly to the two conditions). Dunnette and Heneman (1956) present data showing more favorable attitude-scale responses under high-threat conditions (the personnel manager of the respondents' company administered the questionnaire) than under low-threat conditions (a research team from a university served as survey administrator). Neither of the groups in the Dunnette and Heneman study were identified and participants were explicitly asked not to sign their names to their questionnaires.

These two studies suggest that more than just the issues of identification versus anonymity are important potential sources of distortion in industrial opinion-survey research. It would appear that implicit threat resulting from the environment in which survey forms are filled out tends to influence the degree of response favorability within industrial settings. One important variable in the administration of opinion surveys not covered in these studies is the question of possible differences in the data resulting from mailed versus group-administered opinion questionnaires.

Only two studies were found, neither in an industrial setting, which dealt specifically with this aspect of survey administration. Barnett (1965) reports data on attitudes of 202 college students toward different facets of interfaith marriage which suggests more favorable opinions when the data were collected within the classroom than when the questionnaire was completed outside of the classroom. However, the results are somewhat ambiguous, and no statistical analysis is presented. Bell, Hartup, and Crowell (1962) found that mothers who completed a Parental Attitude projective questionnaire at home had significantly lower total scores based on 20 scales than did mothers who participated in a supervised group administration. They hypothesized that mail administration resulted in more deliberation and intellectualization and therefore fewer positive endorsements than is the case in group administration. Neither of these studies, however, provides any real insight into the potential data distortions of industrial opinion surveys which may be attributed to mail- versus group-techniques of questionnaire administration.

This paper deals with two basic issues important in the administration of opinion questionnaires in industry: (*a*) it contrasts opinion data collected through mail administration of questionnaires with data collected through group administration under controlled situations; and (*b*) it contrasts data collected by having the respondent indicate his answers on the questionnaire booklet itself with data collected by means of a separate answer sheet. The purpose of the study was to see whether these different administration conditions yielded comparable results or whether there are any evidences of bias as a result of these manipulations in the method of administration, apart from any explicit consideration of respondent identification or anonymity.

These questions have practical significance, because for each of the two comparison situations there are rather powerful considerations arguing in favor of one approach over the other. Group administration of the industrial questionnaire at the place of work appears more desirable than mailing it to employees' homes because: (*a*) it provides for uniform conditions under which the data are collected; (*b*) it provides an opportunity for the survey administrator to answer questions and clarify any ambiguous points; (*c*) it tends to yield a higher rate of response than a mail survey; (*d*) it permits the data to be collected in a relatively short period of time; and (*e*) there is some saving on the expense of postage. On the other hand, group administration carries the potential danger of inhibiting response through some unknown amount of apprehension about possible lack of anonymity from having the forms completed and collected on

the company premises; there is a possibility of group norms operating in some inexplicable way to influence responses; there is a possibility that employees will feel they were forced into participating against their will.

With regard to the second issue addressed in this study, there are several potential advantages to using an answer sheet to record survey data rather than having respondents answer directly in a questionnaire booklet. Most importantly, answer sheets can be machine scored and converted into punched cards eliminating the time consuming and expensive process of key-punching. Answer sheets are also easy to handle and maintain. In addition, the questionnaire booklets may be utilized more than once. On the other hand, there is probably more potential for respondent error in recording responses on an answer sheet than exists when the actual response alternative is circled in a questionnaire booklet.

For this study, two hypotheses were formulated with regard to these issues:

Hypothesis 1: Employees who complete an opinion survey under group-administration conditions will register more favorable attitudes in certain specific attitude areas than will a comparable group of employees participating in the survey through mail-administration conditions. Specifically, it was hypothesized that ratings of immediate management, of work associates, and of salary would be higher under the office-administration condition than in a mail survey; it was presumed that group norms would tend to yield relatively high ratings of work associates, and a reluctance to criticize local management—either directly or in terms of the manager's salary-administration activities—would influence opinions in these attitude areas.

Hypothesis 2: There will be more error in the data collected through use of an answer sheet to record survey responses than will be true in data collected by circling responses within the questionnaire booklet. Operationally, it was hypothesized that the variance of the distributions of responses collected by answer sheets would be inflated by a larger component of error variance than would the distributions for data collected under the booklet conditions, and thus the answer sheet condition would exhibit larger total variance.

METHOD

Data for the study were gathered from responses to a 150-item opinion questionnaire administered to a population of 13,000 technicians in a large national organization. The population of technicians was physically located at roughly 200 different locations. A representative sample of 16 of these locations was selected to be administered the questionnaire in small groups within the office. In the group sessions it was explained that the results were completely anonymous and participation was voluntary (several individuals in fact elected not to complete the survey). In the group session, each individual was given a separate envelope into which he could place his completed questionnaire before they were collected and mailed to the survey research representative.

There were 843 employees in these 16 specially sampled offices. Approximately half of them were instructed to circle their responses to the questions directly within the questionnaire booklet and return the entire booklet; the other half blackened the space on a standard test-scoring answer sheet which corresponded to their response to the particular questionnaire item. The answer sheets contained spaces for all 150 items on one side of an $8\frac{1}{2} \times 11$-in. sheet of paper and were designed so that they could be read directly into an IBM 1232 optical mark scoring reader producing punched-card output.

The remainder of the population of technicians received the questionnaire in the mail at their homes along with a stamped return envelope addressed directly to the survey researcher. All of these individuals utilized the optical mark scoring answer sheet. The mail survey resulted in a 75% response.

A second sample of 16 offices was selected from the mail-administered portion of the survey to match the "in-office" group of 16 offices as closely as possible in terms of size (number of employees assigned to the office) and in terms of geographic area. Data from this matched sample of 16 offices form the mail-administered condition discussed in this study; there were 979 technicians included in this sample.

To summarize, there were three conditions of survey administration which are contrasted in this paper:

(*a*) a group of 357 technicians who circled their responses to the questions within the questionnaire booklet and who completed the questionnaire in the office under group conditions (*Office, Booklet Condition*);

(*b*) a group of 486 technicians who utilized the answer sheet, completing the questionnaire in the office (*Office, Answer Sheet Condition*); and

(*c*) a group of 979 technicians, utilizing the answer sheet, who responded to the survey at home following mail administration (*Home, Answer Sheet Condition*).

RESULTS

Responses to 41 key questionnaire items which asked the respondent to rate his degree of satisfaction or dissatisfaction with various aspects of his work role were factor analyzed. Eight distinct components being measured by the survey were identified following an oblique rotation of the data. These factors were:

1. Relationship with immediate manager.
2. Attitude toward the company in general.
3. Perceived demands of the job.
4. Rating of the work itself.
5. Attitudes regarding future opportunities.
6. Extent of harmony and cooperation in the work environment.
7. Feelings of job security.
8. Evaluation of salary.

The highest loading items within each of these eight factor areas were chosen for analysis in testing the two hypotheses. Altogether, 32 items were chosen with factor loadings ranging from .36 to .78.

To investigate the effects of type of administration on response—the first hypothesis—t tests of the difference between the mean scores of the selected 32 items were computed for each of the three questionnaire-administration conditions. As Table 1 indicates, comparison of the scores between the two groups that responded in the office shows that only 2 of the 32 items were answered significantly different at the .05 level of significance (two-tailed tests). This result is what one would expect by pure chance alone. However, when the data for the group that participated in the mail survey are contrasted with those of either of the in-office groups (both Booklet and Answer Sheet groups) a larger than expected number of significant t's result. There are six significant t's at the .05 level (one-tailed tests) for the Office, Booklet versus Home, Answer Sheet group ($\chi^2 = 8.53$, $p < .005$), and 11 for the Office, Answer Sheet versus Home, Answer Sheet group ($\chi^2 = 43.20$, $p < .001$). As hypothesized, for a significant number of items, the Office group responded more favorably than did the group which participated in the mail survey. The data suggest that the largest frequency of significantly different means occurs in the attitude component dealing with "the company in general." The major items reflecting these differences were:

How would you rate the XYZ company as a company to work for compared with other companies you know about?
If you knew when you joined the company what you know now, would you come to work here?
Do you feel that the top managers of XYZ company have a genuine interest in the welfare and happiness of company employees?

The remaining significant t's tend to be scattered through the other attitude components. Thus, the basic part of the first hypothesis was confirmed, but the subhypothesis regarding the specific attitude areas where differences would be most marked was not confirmed.

To evaluate the second hypothesis, that is, that error would be introduced by having *S*s respond on a separate answer sheet, F tests were conducted on the variances of the 32 selected items. A pattern similar to that found from the t tests was evident. Between the two In-Office groups (Office, Answer Sheet versus Office, Booklet) only one pair of item variances differed significantly at the .05 level. But, between each of the In-Office groups and the Mail group, there were seven significant F ratios ($\chi^2 = 13.33$, $p < .001$). In all

TABLE 1

NUMBER OF QUESTIONNAIRE ITEM COMPARISONS EXCEEDING THE INDICATED LEVEL OF SIGNIFICANCE FOR t TESTS AND F TESTS COMPARING THREE DIFFERENT SURVEY ADMINISTRATION CONDITIONS FOR 32 SELECTED ITEMS

Administration conditions compared					
Office, Booklet versus Office, Answer Sheet		Office, Booklet versus Home, Answer Sheet		Office, Answer Sheet versus Home, Answer Sheet	
t	F	t	F	t	F
1***	0***	3***	3***	5***	3***
2**	1**	6**	7**	11**	7**
5*	3*	12*	9*	16*	10*

* $p < .10$.
** $p < .05$.
*** $p < .01$.

cases, the variance of the item when answered by the Mail group was greater than the variance of the item when answered by the In-Office group. Again, a preponderance of the significant F ratios was found in the items clustered under the factor of Company in General and the remainder scattered among the other attitude components. Thus, the second hypothesis dealing with differences in answer-sheet versus booklet response modes was not confirmed by the data.

DISCUSSION

It would seem that a certain amount of distortion occurs when employees respond to an opinion survey in group sessions at their place of work as opposed to responding in some other setting. These results suggest that, with a greater than chance frequency, employees tend to give a more favorable response to opinion questionnaires when surveyed at their office locations than when surveyed by mail at home. The tendency is particularly true with regard to global evaluations of the company as a place to work, which tends to parallel part of the findings of Klein et al. (1967) regarding attitude areas most susceptible to distortion under threat.

These differences have practical significance for the interpretation of opinion-survey results as well as statistical significance. Despite the large sample sizes utilized in this study, from 5 to 8% fewer individuals responded to the favorable response alternatives under the mail conditions than under the in-office condition for those items yielding significant t tests. Distortions of this magnitude are clearly important in the interpretation of opinion-survey data.

No definite evidence can be offered from the data at hand as to which is the "true" response, nor is it at all clear why this tendency is particularly evident with items dealing with the company in general. Perhaps the fact of assembling groups of employees and taking time away from the workplace to obtain their opinions about a number of important areas serves as a tangible symbol of company concern for the individual and tends to influence ratings of the company in a favorable direction—a situation analogous to a "Hawthorne effect" spilling onto the company in general. Conversely, it may be that a survey received in the mail is not accompanied by this same type of tangible evidence of concern.

An alternate explanation might assume that there is a higher probability of inhibitions to honest and frank response in the within-office condition than under the mail condition; presumably, this type of inhibition, whether conscious or unconscious, would tend to inflate the level of favorability of survey data collected in group sessions but not to operate in the more demonstrably anonymous mail survey.

A question dealing with the respondent's preferred method of survey participation was included among the 150 general-opinion questions in the questionnaire. Responses to this question show that a majority (62%) of these technicians preferred to fill out the questionnaire at their convenience and mail it back. Those who received the questionnaire in the mail overwhelmingly preferred the mail method (86% preferred it); a significant number of those who filled it out in a group meeting preferred the group method (44%), but a significant number of the group participants (33%) would have preferred instead to have received it in the mail at home; the remaining group participants (23%) stated they would have preferred to have the questionnaires handed out at work to be completed at their convenience and mailed in directly. This majority preference for non-group-survey administration tends to suggest that inhibitions operating in the group situation may be at least partially responsible for the differences observed in this study. However, at the present time the exact reasons for the observed differences must remain a matter of conjecture; the significant points are: (*a*) that the two administration conditions do not yield completely comparable results, and (*b*) that the in-office administration conditions tend to yield more favorable ratings than mail administration.

The second hypothesis dealing with errors introduced through use of the answer sheet was not supported. The lack of differences between the two groups which answered the questionnaire in the office tends to indicate that, as such, the separate answer sheet had no

effect on the respondents. If, as was hypothesized, the separate answer sheet had served to increase the number of errors in the mechanics of answering, one would expect the item variances for that group to be larger than the item variances for the group not using the answer sheet.

However, the results do lend support to the idea of distortion caused by the method of administration. For a significantly greater than expected number of comparisons, the item variances of the mail-administration group exceeded those of the office-administration conditions. Perhaps this increased variance results from a reduction in pressure (either conscious or unconscious) to respond in a socially acceptable fashion, or perhaps respondents feel less constraint against using the full scale of the response alternatives when answering within the relatively neutral atmosphere associated with the mail survey. We are unable to say what the reasons are from these results. But the data do clearly suggest significant differences between mail- and group-survey administration conditions.

REFERENCES

BARNETT, L. D. Responses to questionnaires completed inside and outside the classroom. *Family Life Coordinator*, 1965, 14, 130–132.

BELL, R. Q., HARTUP, W. W., & CROWELL, D. H. Mailed versus supervised administration of a projective questionnaire. *Journal of Consulting Psychology*, 1962, 26, 290.

DUNNETTE, M. D., & HENEMAN, H. G., JR. Influence of scale administrator on employee attitude responses. *Journal of Applied Psychology*, 1956, 40, 73–77.

KLEIN, S. M., MAHER, J. R., & DUNNINGTON, R. A. Differences between identified and anonymous subjects in responding to an industrial opinion survey. *Journal of Applied Psychology*, 1967, 51, 152–160.

PELZ, D. The influence of anonymity on expressed attitudes. *Human Organization*, 1959, 18, 88–91.

ROSEN, N. A. Anonymity and attitude measurement. *Public Opinion Quarterly*, 1960, 24, 675–679.

(Received December 7, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 503-508

AN ENQUIRY INTO THE EFFECT OF EXPOSURE TO ADVERTISEMENTS ON SUBSEQUENT PERCEPTION OF SIMILAR ADVERTISEMENTS

DALE H. BERG

University of Alberta

This study tests the hypothesis that forced exposure to advertising lowers the threshold for perceiving the advertisements. The enquiry gains its import in its relevance to the basic question: what is the effect of continued exposure? Using 2 groups, Control and Experimental, of late-teenage girls, a straightforward "before-after" experiment was designed wherein the experimental group was exposed to a certain form of advertising. The study consisted of 3 phases. The 1st phase involved the determination of a "perceptual sensitivity" base score. The 2nd phase involved exposing the Experimental group to the form of advertising under study. The 3rd phase consisted of remeasuring the perceptual sensitivity of both groups. The study demonstrated that forced exposure does increase perceptivity. The Experimental group showed significant differences in perceptual level from the Control group when remeasured in the 3rd phase.

The present century has witnessed tremendous changes in the use of mass media to sell commodities and to control people. The utilization of mass communication has expanded at almost geometric progression. Placed in the right magazines or newspapers, televised at the right time, or broadcast via radio, the desired message can reach the eyes or ears of a small nation of people.

In spite of the fantastic possibilities of exposing populations to propaganda, questions persist regarding the effects of exposure. It is, for example, one matter to say that 800,000 readers have seen a particular issue of a magazine, and an entirely different matter to contend that through their exposure they have absorbed the desired message.

The faith placed in the effectiveness of the mass media is evidenced by the amount expended each year on advertising and advertising research. In spite of the fact that a few isolated cases exist in which a company has not advertised and has been successful (e.g., The Hershey Company), the mass media are regarded as a highly effective means of changing attitudes and consumer habits.

Some investigators in the field have suggested that the impact of advertising is gained through repetition. The idea is that perception of advertisements is generally at such a low level of interest that it takes a considerable amount of exposure to produce even a negligible effect. It has been suggested that much advertising material is learning according to principles that hold for the memorization of meaningless nonsense material (Krugman, 1965).

Lucas reports that when readers of magazines were shown current issues and asked to report on advertisements that were familiar to them, many reported familiarity with ads which had not yet been published (Lucas, 1940). This finding would indicate that familiarity is at best superficial.

In spite of the suggested low level of awareness of magazine advertisements, reader response indicates that at least some people are concerned about the content of advertisements. Recently, there has been a considerable expression of concern regarding advertisements of brassieres and girdles that picture models in these garments. Frequently, distraught readers of the various publications that carry these advertisements write letters to the editor expressing fears that these photographs are having adverse effects on those who see them, especially males and children.[1]

In spite of Krugman's (1965) contention that much of advertising material is learned

[1] Letters to syndicated "lovelorn" columnists, and "call-in" radio programs, women's magazines, newspapers.

TABLE 1

AGE DISTRIBUTION OF SUBJECTS

Age	Control group	Experimental group
16	1	—
17	5	7
18	18	16
19	6	6
20	—	1

on a very low level of interest, the amount of resistance stimulated by using photographs in brassiere and girdle advertising suggests that such advertisements are being noticed on more than a superficial level by at least some viewers. Krugman's conclusions might still be applied to the copy of these advertisements less often noticed.

A key issue remains the question of the effects of continued exposure. Shall it be assumed, for instance, that continued exposure results in increased awareness? It would then seem that by the continued use of such advertising it may be possible to reduce and perhaps even eliminate public aversion. It might be postulated that increased perceptivity reduces aversion to such forms of advertising, if for no other reason than that of familiarity.

At least two questions are raised: (*a*) Does exposure to advertising lower the threshold for perceiving the ads? and (*b*) If so, is increased sensitivity associated with an altered meaning assigned to the stimuli (e.g., reduced aversion)?

This investigation is concerned with the first of these questions.

This study tests the hypothesis that forced exposure results in increased perceptivity. Using late-teenage girls as *Ss*, an attempt was made to increase their awareness of brassiere and girdle advertisements by experimentally exposing them to a considerable amount of such advertising.

PROCEDURE

Subjects

Sixty young women residents in the first-year class of a nursing school served as *Ss*. These 60 were drawn from a class of 117 girls. These women ranged in age from 16 to 20, with the mean age at 18. The modal age (34 *Ss*) was also 18. These 60 women were divided into an Experimental and a Control group with age distributions as shown in Table 1.

All participants in the study, even those eliminated in the first phase, were purposely misled regarding the goals of the study. Prior to any testing, the students were told that the study was concerned with their reactions to all forms of magazine advertising. The researchers were said to be interested in their reactions to various layout designs, colors, copy, and brand-name effectiveness. In other words, what was supposedly being studied was their personal reaction to numerous magazine advertisements.

Prior to the second phase, *Ss* were instructed to give the advertisements under study their utmost attention and to respond in terms of their personal reaction. They were impressed with the idea that what was desired was their honest reaction so that advertising effectiveness could be reliably studied.

Follow-up questioning, to be described later, indicated that the deception worked without exception. The *Ss* seemed unaware of the actual purpose of the study.

Technique

A stereoscope is used to present an *S* with two different pictures, one to each eye, simultaneously, for a "just-perceivable" interval.

The present investigation represents a modification of the usual technique, dual projectors with coupled shutters, in that we have used a viewbox with divided perceptual fields. Contrasting pictures were thrown against a translucent screen so that a different image was presented to each eye. Using a single 35-mm. transparency with two pictures on it, positioned to project exactly on the two perceptual fields, it was possible to use only one projector mounted with a shutter. In this way the problem of synchronizing two shutters with a mechanical coupling device was eliminated. The use of one shutter assured that the presentations were simultaneous.

The one problem which was encountered was not a mechanical one. It became obvious at the very beginning that the material which was to be projected, magazine advertisements, was quite complicated compared to the material previously used by researchers using this stereoscopic technique. Advertisements contain not only pictures, but wording of various sizes. Color combinations within the ad also added to the complexity of the stimuli. Thus, it became necessary to adjust the usual length of exposure in such a manner as to make it possible to perceive the pictures being presented and yet retain the "just-perceptible" effect upon which the technique depends. It is essential that *S* have only enough time to perceive the pictures, without actually being cognizant of them. What was needed was a shutter speed which was slow enough to allow an image to be formed on the retina and fast enough to prevent *S* from reporting both pictures or from

cognitively appreciating both pictures for such a period of time as to allow her to choose consciously which picture she would report. Preliminary studies indicated that $\frac{1}{25}$ of a second allowed "time for perception" without "time for recognition."

Color advertisements of all varieties were clipped from magazines of the large size (10×13 in.). These advertisements were paired using several criteria such as color similarity and similarity in terms of writing position and size of photographs. It was felt that color matching would reduce the effect of color dominance especially found in reds. About 50 pairs of advertisements were photographed, each pair being different. Of these, 20 pairs contained a brassiere and/or girdle advertisement using actual models. These ads were paired with non-brassiere-girdle advertisements. The brassiere-girdle advertisements, from here on referred to as the "critical pictures," were alternated from left to right positions.

Alternating the critical pictures assured that the effects of optic dominance would be eliminated. Eleven critical pictures appeared on the right side, the remaining nine on the left.

All of the slides were photographed against a dull black background so that in projection the only light appearing on the viewing screen would be that of the advertisement. The result was that extraneous light and other material were eliminated from the viewing field. A 300-w. projector was used because it produced a bright image without "burning it in" on the retina.

It was decided as a result of the preliminary studies that a "warm-up" series of slides would be presented each *S* just prior to the "actual" run of the 50 slides. Preliminary studies indicated that not only did *Ss* need a warm-up period to accustom them to the device, but that some also required that their "blocks" to saying "brassiere" or "girdle" be broken down. Twenty slides were chosen for the warm-up series, though in most cases the warm-up required no more than 12 slides. In the warm-up series, the eighth and thirteenth slides were used to break down verbal inhibitions and were labeled "forcers."

This preliminary testing was conducted on *Ss* outside the population from which the Control and Experimental groups were selected. Thus contamination of the *Ss* of report has been avoided.

Method

The study may be considered in three phases. Phase One involves the determination of some base against which perceptual change can be measured. Phase Two is the "educational" segment of the study. Over a period of 5 wk., in biweekly sessions, *Ss* were asked to judge the quality of actual advertisements clipped from magazines. This phase is critical in that the Experimental group was exposed to brassiere-girdle advertisements, while the Control group rated "innocuous" ads. Phase Three measures change in perceptual sensitivity.

Phase One. The first phase, like the third phase, consisted of 50 paired projections of advertisements shown at $\frac{1}{25}$ of a second. Of the 50 paired projections, 20 were critical slides (defined earlier). Thus the number of critical slides perceived out of the 20 constitutes the base score in Phase One and the comparison score in Phase Three. Both Experimental and Control groups were run through this series.

Fifty paired projections were used in an effort to conceal the critical slides so that *Ss* would not become aware of the nature of the enquiry. As noted before, concerted effort was made to influence the *Ss* to think that the study was concerned with all forms of advertising. Investigation by random interviews with the *Ss* at the conclusion of the study revealed that the 30 noncritical slides accomplished this desired end.

The *Ss* were scheduled in 15-min. intervals for Phase One, allowing ample time for the warm-up series and the actual test series. At the beginning of Phase One, each *S* was given a series of warm-up slides. This allowed adjustment to the dark and to the procedure.

The *Ss* were told to report all that they saw with both eyes or only one eye if they only saw one picture. They reported to a female stenographer sitting next to them. This insured that *Ss* would not be inhibited by having to report seeing a brassiere or girdle to a male recorder.

Slides Number 8 and 13 were the "forcer slides" in the warm-up series; that is, the "forcer slides" contained pictures of brassieres and girdles. Slide 8 contained only one picture. The left side of the slide was opaque. The right side of the slide contained a picture of a girl in a brassiere. In this manner *S* was forced to see a brassiere and thus forced into the situation of having to report the same. Slide 13 contained pictures of brassieres and girdles in both pictures. This too forced the *S*.

If *S* reported "brassiere," without hesitancy on Slide 8, Slide 13 was omitted. Slide 13 was only used as a "forcer" in the event *S* blocked or hesitated unduly long on Slide 8.

Those *Ss* who were too inhibited, that is, who blocked on both "forcers" even after several exposures to them, were omitted from the study. Those *Ss* who were not able to adjust to the $\frac{1}{25}$ speed after extensive warm-up were also rejected. The number of *Ss* rejected for being too slow or for being overly inhibited was small, that is, 10 and 3, respectively.

An Experimental and a Control group comprised of 30 *Ss* each were selected from *Ss* tested. The two groups were formed through a process of matching on Phase One scores and randomly distributing the paired *Ss* into either the control or experimental category. The number of critical slides perceived by each *S* is reported in Table 2. The total number of critical slides perceived in Phase One by the Control group was 198 as compared to 199 in the Experimental group. The mean number of critical slides perceived is 6.6 for each group.

TABLE 2

NUMBER OF CRITICAL SLIDES PERCEIVED IN PHASES ONE AND THREE BY SUBJECT PAIRS IN THE EXPERIMENTAL AND CONTROL GROUPS

Subject pair number Con/Exp	Age		Phase One/Three scores and amount of change							
			Control group subjects				Experimental group subjects			
			Phase		Change		Phase		Change	
	C	E	I	III	+	−	I	III	+	−
1	18	19	3	8	5		3	13	10	
2	18	17	3	6	3		3	13	10	
3	18	19	4	7	3		4	12	8	
4	17	18	4	4			4	4		
5	19	18	4	13	9		4	6	2	
6	18	18	4	4			4	9	5	
7	19	18	4	5	1		5	10	5	
8	17	18	5	2		3	5	13	8	
9	18	18	5	7	2		5	10	5	
10	18	18	5	6	1		5	7	2	
11	18	18	5	6	1		5	11	6	
12	19	17	5	4		1	5	7	2	
13	18	17	5	8	3		6	13	7	
14	18	19	6	4		2	6	9	3	
15	18	19	6	9	3		6	12	6	
16	18	17	6	9	3		6	11	5	
17	18	18	6	7	1		7	12	5	
18	18	19	7	6		1	7	10	3	
19	16	18	7	4		3	7	7		
20	19	18	7	9	2		7	11	4	
21	17	18	8	9	1		8	12	4	
22	18	18	8	8			8	6		2
23	19	18	8	11	3		8	13	5	
24	17	17	8	5		3	8	17	9	
25	18	19	9	9			9	7		2
26	18	20	10	12	2		10	9		1
27	19	18	10	10			10	6		4
28	18	18	10	9		1	10	13	3	
29	17	17	12	10		2	11	12	1	
30	18	17	15	11		4	12	16	4	
Total			199	222	43	20	198	311	122	9
Means (N = 30)			6.6	7.4	1.4	.67	6.6	10.36	4.0	.3

Phase Two. The second phase of the study consisted of exposing the Experimental group to the critical advertisements while the Control group viewed noncritical advertisements.

Actual advertisements, clipped from magazines, were given to the Experimental and Control groups separately for their judgment as to "quality" and "effectiveness."

These groups met twice a week for 5 wk. and were exposed to 30 advertisements at each meeting. Exposure was limited to 1 min. per advertisement, which time included completion of the "ad judgment" form for each clipping. In total, then, each S was exposed to 300 advertisements.

The Experimental group differed in exposure from the Control group only in that it evaluated 12

critical ads, of the 30 each night. The noncritical ads were identical for both groups. The Experimental group, then, received a 40% "dose" of critical advertisements as the "indoctrination" course.

To control for the problem of postsession intercommunication between the Control and Experimental groups, all *Ss* were reminded at each session not to discuss the test with the other group or between themselves, as their individual judgments were extremely important. The request was apparently well followed as indicated by posttest questioning.

In the event that talking might occur, it was decided to include five critical advertisements in the third session Control group packet. Thus if the Experimental group *Ss* would have remarked that they had been viewing an excessive number of advertisements for brassieres and girdles, they would encounter the reply that the other group also had this experience.

The effectiveness of this deception rested on the assumption that the actual number of critical advertisements given to the Experimental group would not have been counted and therefore the definition of what was excessive would be sufficiently devoid of meaning so as to remain unnoticed.

Phase Three. After all the exposure sessions were completed and the few absences made up, Phase Three, the final phase, was conducted. This retest took place over three evenings. The *Ss* were scheduled every 10 min. for their projection sessions. Only one repeat of any slide was allowed and this was permitted to occur only after extensive quizzing for any possible image by the recorder. The quizzing made it obvious that there was little chance of having a repeated projection. Perception became much better after one encounter with the probing questions, "Didn't you see any image at all? Are you sure?"

Since *Ss* were familiar with the stereoscopic technique they warmed-up quickly and the 10-min. scheduling proved to be quite adequate. Three slides were generally sufficient to adjust the eyes and the "forcer" (Slide 8) was given in fourth position. If *S* responded "brassiere" to Number 8, she was given one more slide for deceptive purposes and the series of 50 slides was begun, as in Phase One. The *Ss* were always told when the actual series began and the warm-up period had ended. As in Phase One, all responses were recorded.

A schematic diagram of the study is included in Table 3. The diagram is divided into three phases and labeled as to the content in each phase. The simplicity of design made the interpretation of the results quite straightforward.

TABLE 3

Schematic Diagram of the Three Phases

Phase One	Phase Two	Phase Three
Ascertainment of differential thresholds for perception of sex-oriented stimuli.	Introduction of experimental variables.	Ascertainment of change in differential thresholds for perception of sex-oriented stimuli.

Results

The results, the number of critical slides perceived in Phase One and Three, are given in Table 3 for both the Control and Experimental groups. The table also indicates amount of change. The numbers assigned to *Ss* are matched as *Ss* were paired, thus *S* 1 in the Experimental group has her counterpart *S*, Number 1, in the Control group.

The mean number of critical slides perceived by both groups in Phase One was 6.6. The mean number of critical slides perceived in Phase Three was 7.4 for the Control group as compared to 10.36 for the Experimental group. In terms of number of critical slides perceived, this means that the Experimental group perceived 113 more critical slides (net) in Phase Three as compared to the 23 more critical slides (net) by the Control group.

The null hypotheiss that the experimental variable, exposure, had no effect was assumed and tested at the .05 level of significance with a one-tailed test.[2]

With 29 degrees of freedom, if $t = 1.699$, the null hypothesis would be rejected at this (.05) significance level. The *SD* was calculated to be 3.45, and t to be 4.62. The null hypothesis was rejected.

Discussion

A classic definition of "attitude" holds this term to refer to "a tendency to respond to certain stimuli in a particular manner" (Trasler, 1962). The present study demonstrates that forced exposure to stimuli (familiarity) lowers the perceptual limen. This is one measure of attitude change, and it gains significance here because of our use of actual advertising materials, rather than less complex stimuli.

As in all studies of attitude, the meaning of this particular form of attitudinal determi-

[2] A one-tailed test was used because direction of change had been predicted.

nation for other dimensions of behavior remains at issue.

From this and similar investigations it may be assumed, for example, that the advertiser can increase "product awareness" through repetition. But the relationship between such increased perceptivity of the advertiser's message and action upon it (buying) is a continuing question.

Further, the findings intrude upon the censorship debate. It is clear that exposure to sometimes offensive materials increases the perception of them. What needs to be known is whether such familiarity reduces or increases the emotional valence of these stimuli. That is, does viewing items of secondary sexual content reduce their impact, or does it stimulate other dimensions of attitude and behavior?

REFERENCES

KRUGMAN, H. E. The impact of television advertising: Learning without involvement. *Public Opinion Quarterly*, 1965, 29, 349–356.

LUCAS, D. B. A rigid technique for measuring the impression values of specific magazine advertisements. *Journal of Applied Psychology*, 1940, 24, 778–790.

TRASLER, G. *The explanation of criminality*. London: Routledge & Kegan Paul, 1962.

(Received December 8, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 509–520

CONVERGENT AND DISCRIMINANT VALIDATION OF SATISFACTION AND DESIRE MEASURES BY INTERVIEWS AND QUESTIONNAIRES

CLAYTON P. ALDERFER [1]

Graduate School of Business and Public Administration, Cornell University

The problem under study was to examine the convergent and discriminant validity of interview and questionnaire measures of satisfaction and desire in 5 different need categories. Using the Campbell-Fiske multitrait-multimethod framework, the results showed relatively high discriminant and convergent validity for the satisfaction scales, low convergent and moderate discriminant validity for the desire scales. The results suggest that the interview versus questionnaire controversy is not necessary; they also suggest that further work is needed in developing desire scales.

Interviews and questionnaires are frequently viewed as alternative sources of data for many problems in behavioral science research. If one method is used, very often the other is not. Interviews have been criticized because of their potential for having the interviewer draw from the respondent what he wants to hear (Dunnette, 1962, pp. 302–303; Hyman & Sheatsley, 1954, pp. 50–122). Questionnaires, on the other hand, have been criticized because their impact on respondents often tends to be viewed as agonizing repetition of trivial questions in the misguided quest of science (Hechinger, 1966). The use of interviews is based on the assumption that effective tactics create an emotionally satisfying setting where the respondent is free to give his own views and where the investigator allows himself to be open to more information than is available in the usual fixed-response questionnaire. A questionnaire, however, is more amenable to precise quantification through scaling techniques.

Until recently there has been little systematic comparison of the two measurement techniques. Sears (1965), however, has shown that mothers' attitudes toward permitting sexual and aggressive behavior in their children tend to be highly correlated when measured by interview and then by questionnaire. The purpose of this paper is to present the results of a study designed to assess the validity of interview and questionnaire measures of need satisfaction and desires by the multitrait-multimethod procedure (Campbell & Fiske, 1959).

Theoretical Constructs

The variables measured by interview and questionnaire were based on a conceptual scheme consisting of three need categories: *existence, relatedness,* and *growth.* For each need category, two attributes were measured. The first was desire. How important was a particular reward? How much did the respondent want it? The second was satisfaction. Satisfaction for each group of needs is defined in terms of a particular end state and a process for attaining that state. Existence needs have physical or material end points; they are reached by individuals getting absolutely or relatively enough. Pay and fringe benefits are the particular existence needs considered in this study. Relatedness end points refer to emotional states with regard to specific other people; the process for attaining them consists of two or more people sharing their thoughts and feelings. Respect by Superiors and Respect by Peers are the two relatedness needs of interest in this study. The end points of growth needs refer to the solutions of problems which lie within the actual or potential capacities of

[1] The material in this paper is adapted from an unpublished doctoral dissertation in the Department of Industrial Administration at Yale University, 1966. The writer would like to express his appreciation to Charleen Alderfer and to Chris Argyris, Edward Lawler, Vernon Buck, Thomas Lodahl, and Tim Hall for their help in the project.

TABLE 1

SATISFACTION ITEMS FACTOR ANALYSIS

Item	Varimax factors: Pay	Fringe Benefits	Respect by Superiors	Respect by Peers	Growth
1. Considering the work required, the pay for this job is fair. (Agree)	*.74*	.18	−.01	.02	.04
2. Compared to the rates for similar work here, my pay is good. (Agree)	*.82*	.05	.11	.04	.07
3. My pay is adequate to provide for the basic things in life. (Agree)	*.37*	.06	−.02	.09	.15
4. In general I am satisfied with my pay. (Agree)	*.77*	.16	.17	.04	.15
5. Compared to the rates for similar jobs in other places, my pay is poor. (Disagree)	*.76*	.15	.09	.12	−.04
6. I find it hard to do some of the things I enjoy most on the pay I receive. (Disagree)	*.56*	.21	.11	−.03	.13
7. Compared to the rates for less demanding jobs, the pay for my job is poor. (Disagree)	*.77*	.16	.06	.02	.05
8. In general I am satisfied with the fringe benefits I receive. (Agree)	.19	*.58*	.06	.09	.02
9. My job offers good fringe benefits. (Agree)	.14	*.46*	.14	.07	.06
10. Compared to other places our fringe benefits are excellent. (Agree)	.28	*.63*	.00	.06	.05
11. The fringe benefit program does not give the security I want. (Disagree)	.22	*.70*	.10	.03	.08
12. The fringe benefit program here needs improvement. (Disagree)	.16	*.79*	.05	.00	.04
13. Our fringe benefits do not cover many of the areas they should. (Disagree)	.19	*.73*	.19	.15	.15
14. It's easy to talk with my boss about my job. (Agree)	.04	.07	*.75*	.13	.03
15. My boss takes account of my wishes and desires. (Agree)	.10	.10	*.68*	.09	.14
16. My boss keeps me informed about what is happening in the company. (Agree)	.06	.10	*.67*	.18	.07
17. My boss often asks for my opinion about the job. (Agree)	.04	.03	*.59*	.18	.37
18. My boss gives me credit when I do good work. (Agree)	.11	−.05	*.71*	.05	.21
19. My boss keeps a close check on his people. (Disagree)	−.11	.01	*−.37*	.06	.03
20. My boss will play one person against another. (Disagree)	.00	.05	*.67*	.06	.06
21. I feel that there are certain "right" things to tell my boss. (Disagree)	.07	.03	−.06	−.05	−.04
22. My boss does not let me know when I could improve my performance. (Disagree)	.25	.18	*.61*	.10	.02
23. My boss expects people to do things his way. (Disagree)	.05	.00	*.40*	.17	.09
24. My co-workers respect my opinions when I disagree with them. (Agree)	−.07	−.11	.07	*.57*	.19
25. I can count on my co-workers to give me a hand when I need it. (Agree)	.05	.04	.10	*.63*	.08
26. I can speak my mind to my co-workers. (Agree)	.07	.04	−.04	*.67*	.00
27. My co-workers welcome opinions different from their own. (Agree)	.04	.04	.25	*.65*	.15
28. My co-workers are highly critical of each other when mistakes are made. (Disagree)	.05	.02	.03	*.28*	.03
29. My co-workers are uncooperative unless it's to their advantage. (Disagree)	.16	.06	.20	*.68*	.11
30. My co-workers will not stick out their necks for me. (Disagree)	.03	.19	.21	*.62*	.18
31. I have few close friends among my co-workers. (Disagree)	−.06	.23	.10	*.44*	.01
32. I have an opportunity to use many of my skills at work. (Agree)	.16	.04	.20	.15	*.72*

TABLE 1—(*Continued*)

Item	Varimax factors				
	Pay	Fringe Benefits	Respect by Superiors	Respect by Peers	Growth
33. Careful planning is necessary in my job. (Agree)	.09	−.01	.06	.03	*.70*
34. My job requires that a person use a wide range of abilities. (Agree)	.04	−.07	.00	.08	*.79*
35. Few things in my job give me a sense of accomplishment. (Disagree)	−.07	.16	.23	.02	*.42*
36. My major abilities are seldom needed at work. (Disagree)	.06	.14	.26	.05	*.59*
37. In my job I have the same things to do over and over. (Disagree)	.14	.11	.04	.10	*.62*
38. This job has helped me to see some talents I never knew I had. (Agree)	.01	.14	.08	.18	*.65*
39. My job is constantly changing. (Agree)	−.02	.01	.00	.09	*.68*
40. My job requires making one or more important decisions every day. (Agree)	−.06	.00	.03	.12	*.65*
41. I am rarely asked to try new ways of doing things at work. (Disagree)	.08	.03	.07	.16	*.48*
42. I seldom get the feeling of learning new things from my work. (Disagree)	.08	.10	.22	−.09	*.68*
43. I do not have the opportunity to do challenging things at work. (Disagree)	.10	.15	.15	.12	*.68*

Note.—The highest loadings on each factor which are used to define the factor are italicized.

the person. Integrating and/or differentiating one's talents are the processes for satisfying growth needs. Using and learning skills and abilities are the growth needs of interest in this investigation.

Validation Framework

Campbell and Fiske (1959) have presented a systematic procedure for validation which is based upon convergent and discriminant validation of more than one trait by more than one method. The convergent-validity question asks whether different methods purporting to measure the same trait correlate highly with each other. The discriminant-validity question asks whether measures of traits which are purportedly different from each other are uncorrelated.

When the Campbell-Fiske procedure is used, validity information is contained in the "multitrait-multimethod matrix." Each entry in the matrix is defined by the coincidence of one or two trait measurements by one or two methods. In the present study the methods are interviews and questionnaires; the traits are different need categories. (See Table 4.)

The diagonals of the matrix contain reliability and validity coefficients. Here reliability is defined as "the agreement between two efforts to measure the same trait through maximally similar methods." Validity is defined as "the agreement between two efforts to measure the same trait through different methods" (Campbell & Fiske, 1959).

In terms of the matrix, the question of convergent validity has a single criterion. Are the entries on the validity diagonal significantly different from zero and reasonably large? There are three questions to answer for discriminant validity. Is an entry in the validity diagonal higher than the values lying in its column and row where neither trait nor method are common? Is an entry in the validity diagonal higher than correlations between that and other traits using the same method? Is the pattern of correlations among traits the same for each method and for the combination among methods?

METHOD

Subjects

Approximately 300 persons served as *Ss* for the study. They were employees in the largest division

of a manufacturing organization and were randomly selected from a total group of about 1,700 people. The sample was stratified by departments and included job holders from the assembly line through four levels of management. From the total sample, about 130 people were randomly selected to be interviewed. All Ss were asked to complete questionnaires.

Procedure

After an interview was completed a respondent was given a questionnaire and asked to mail it back to the writer at Yale University. The remainder of the questionnaires were given out during group meetings where 10–15 people from different departments were assembled. Depending on the time available the questionnaires were either completed at the meeting or returned to the writer by mail. Approximately 90% of the people who participated in the study returned their questionnaires.

While the interview explanation proceeded, a small tape recorder was setting on a nearby table. The case was open so that it could be clearly seen that it was not turned on. When the initial introduction had been completed the interviewer turned to the recorder and asked the respondent's permission to tape the interview. The person was again reassured that no one in the company would ever be allowed to listen to the tapes. Of the people who came to interviews, 95% gave their permission to record the material.

All of the interviews were conducted by the writer. The tapes were then transcribed verbatim. These transcripts became the raw data for systematic content analysis.

Almost without exception the questions were asked in the same order. They were open-ended, such as "What do you like to get out of a job?" or "Would you describe your boss for me?" There were 15 standard interview questions; the interviews lasted from 20 min. to an hour and a half with a median time of about 35 min. The questions and the exact procedure for content analysis are given elsewhere (Alderfer, 1966). The writer and one other coder scored 75 interviews independently.

Two different questionnaires were used in the study. Each questionnaire was designed to measure satisfaction with, and desire for, two needs in each of the existence, relatedness, and growth areas. They differed, in that one utilized the paired-comparison method and the other a graphic rating scale to assess the desire measures. Satisfaction items were exactly the same on both instruments.

Items for the satisfaction scales appeared in a 7-point Likert-scale format. They were counterbalanced so that in approximately half of the items a person had to agree to be scored for satisfaction and in the other half, disagree to be scored for need satisfaction. For the paired-comparison desire measures, the items appeared two at a time; S was asked which was more important to him. The format departed slightly from the usual paired-comparison procedure in that Ss had the option of saying that neither one was more important than the other. Each pair of items appeared twice. A desire score was obtained by counting the number of times an S selected a given item. The rating-scale measure of desire had each item followed by two ratings, each on a 5-point scale. The first rating was importance; the second was how much more the person wanted.

RESULTS

Satisfaction Scales

The 43 items which were expected to constitute the six satisfaction scales were intercorrelated and factor analyzed by the principal components method. All factors with roots greater than 1 were rotated by the Varimax procedure. The first five factors were clearly interpretable. They can be named Pay, Fringe Benefits, Respect by Superiors, Respect by Peers, and Growth Satisfactions, respectively.

Table 1 gives the items and their loadings on the satisfaction factors. Agree or disagree after the item indicates the direction of the item scored for satisfaction. Because the items were expected to form factors, they are presented together to show which factors they were intended to develop. They did not appear this way in the questionnaires. In general one can say that the factorial structure is quite close to the predictions. The only major exception is that instead of having two growth factors, one for use skills and abilities and one for learn skills and abilities, there is just one. Three minor exceptions were also present. Keeping a close check on people, Item 19, was expected to be seen as lack of trust by the boss, but it loaded in the opposite direction, although the loading was the smallest of the Respect by Superiors items. Telling "right" things to the boss, designed to tap an aspect of perceived conformity, had essentially no loading on the factor. Perhaps this indicates that the item was not consistently understood as intended because Item 23, where the boss expects people to do things his way, does load on the factor. Item 28, having to do with criticism by co-workers, loaded in the direction expected on the Respect by Peers factor, but it did not have a high loading. The item was intended to reflect punitive pressures against risk taking,

TABLE 2

QUESTIONNAIRE SATISFACTION SCALES: NUMBER OF ITEMS AND MEDIAN ITEM CORRELATION

Factor	No. items	r
Pay	7	.50
Fringe Benefits	6	.40
Respect by Superiors	8	.40
Respect by Peers	8	.30
Growth	12	.41

Note.—N = 302.

TABLE 3

INTERVIEW SATISFACTION SCALES: INTERCODER CORRELATIONS

Factor	r
Pay	.81
Fringe Benefits	.77
Respect by Superiors	.84
Respect by Peers	.70
Growth	.76

Note.—N = 75.

but perhaps it was sometimes confused with simply pointing out mistakes.

To estimate the reliability of the factor scores, the median correlation of items defining the factor was used in the Spearman-Brown formula. The number of items is used for the number of replications.[2] Table 2 gives the median correlation and the number of items for each satisfaction scale.[3]

Table 3 gives the raw correlations between the independent interview codings for the five scales. When the coding was begun, the results of the factor analysis were not known. Consequently, separate scores were developed for satisfaction of use skills and abilities and for satisfaction of learn skills and abilities. After the factor-analysis results were available the two scales were added to form the interview growth satisfaction scale. The correlation between the two scales was .49, which is higher than the median item correlations from the questionnaire.

Table 4 presents the multitrait-multimethod matrix for satisfaction scales. Each trait is abbreviated by the letters of its main words. The reliability estimates on the questionnaire-questionnaire and on the interview-interview matrices were calculated by applying the Spearman-Brown formula to the data in Tables 2 and 3. The validity diagonal in the interview-questionnaire matrix comes from the correlations between the interview and questionnaire scale scores for the 111 interviewees who returned usable questionnaires.

Viewing the reliability diagonals one can see that all the values are equal to or greater than .79. The median reliability is .87. The validity diagonal, located in the interview-questionnaire matrix, meets the criterion for convergent validity in four of five cases. All of the entries in the diagonal are significant beyond the .005 level. Only the coefficient for Satisfaction of Respect by Peers is relatively small, and that is larger than any off-diagonal value. All the entries meet the first criterion of discriminant validity: each is higher than any other value in the same row or column. All the entries also meet the second criterion of discriminant validity: each is higher than any correlation between two traits when a common method is used. The third criterion of discriminant validity is met if one is willing to assume that all off-diagonal entries approach zero. Factor-analytic techniques "try" to find orthogonal factors, but the low correlations among interview variables exist without such mathematical gymnastics. Otherwise, there is no immediately discernible pattern among these entries.[4]

Desire Scales

As mentioned above, two different questionnaire procedures were used to obtain desire

[2] The Spearman-Brown formula is $r_s = \frac{mr}{1+(m-1)r}$ where m is the number of replications. In this study, m is the number of items which define the factor; r is the median correlation among items.

[3] Factor scores were computed by the formulas for rotated principal components given by Kaiser (1962).

[4] The correlations in the questionnaire-questionnaire off diagonals of Tables 4 and 8 capitalize on the specific and error components in the data to obtain such nearly perfect orthogonal factors. It would be a worthy task of future research to use the factor-score coefficients reported in this paper on another sample of people to cross-validate the findings reported here (Glass & Taylor, 1966). It is almost certain that the factors would not be as perfectly orthogonal as shown here.

measures. One was a graphic-rating-scale procedure, and the other utilized the method of paired comparisons. Approximately half of the interviewed *Ss* were asked to use each type of questionnaire. Data will be presented concerning the convergent and discriminant validity of each kind of questionnaire with the interview.

The 28 rating-scale items which were expected to constitute the desire scales were intercorrelated and factor analyzed by the principal components method. All factors with roots greater than 1 were rotated by the Varimax procedure. Five of the first six factors were clearly interpretable. They are the same as the satisfaction factors: Pay, Fringe Benefits, Respect by Superiors, Respect by Peers, and Growth Desires.

Table 5 presents the items and their loadings on the desire factors. "Important" or "more" after the item indicates the format in which the variable appeared in the questionnaire. Even though all items appeared in both an "important" and a "more" format, the factor loadings show a different pattern for these two kinds of variable depending on the need area. Existence and growth needs—Pay, Fringe Benefits, and Growth—have high loadings from both types of items on their respective factors. Relatedness needs—Respect by Superiors and Respect by Peers—have high loadings on their "more" items only. Apparently there is not a linear correlation between saying one wants more respect or trust and saying that it is important. This is a decidedly different picture than one obtains in the other need areas and emphasizes the point that one has somewhat different measurement problems with relatedness needs than with others. Items 15 and 16, dealing with using one's approach at work, turned out to load with the Desire for Respect by Superiors items rather than with the Growth items as originally expected. They have the relatedness loading pattern, too, of high on the "more" and low on the "important"

TABLE 4

SATISFACTION SCALE MULTITRAIT-MULTIMETHOD MATRIX

	Questionnaire (n = 302)					Interview				
	P	FB	RS	RP	G	P	FB	RS	RP	G
Questionnaire										
P	(.88)									
FB	.00	(.80)								
RS	.01	.00	(.84)							
RP	−.02	.00	.00	(.79)						
G	−.01	.00	.02	−.04	(.88)					
	(n = 111)					(n = 75)				
Interview										
P	.54	.19	.15	−.18	.17	(.90)				
FB	.13	.47	.11	−.12	.09	.23	(.90)			
RS	.22	−.04	.66	−.13	.21	.19	.10	(.91)		
RP	.13	−.10	.14	.24	.15	.00	.07	.16	(.84)	
G	.19	12	02	.05	.61	.15	.11	.17	.10	(.87)

Note.—The validity diagonal is indicated by the circled coefficients, the reliability diagonals by the entries in parentheses. Constraints of time and money did not allow all the interviews to be scored by both coders. The diagonals of the interview sub matrix are based on n = 75, the number of interviews done by both coders. All of the other interview cells are based on the coder who did all the interviews, n = 111.
Abbreviations. P = Pay, FB = Fringe Benefits; RS = Respect by Superiors; RP = Respect by Peers; G = Growth.

TABLE 5

RATING-SCALE DESIRE ITEMS FACTOR ANALYSIS

Item	Varimax factors				
	Pay	Fringe Benefits	Respect by Superiors	Respect by Peers	Growth
1. Good pay for my work—important	*.67*	.15	−.07	−.02	.08
2. Good pay for my work—more	*.64*	.18	.19	.21	.24
3. Frequent raises in pay—important	*.68*	.33	.06	−.16	−.01
4. Frequent raises in pay—more	*.72*	.08	.44	.03	.10
5. A complete fringe benefit program—important	.25	*.72*	−.08	−.04	.01
6. A complete fringe benefit program—more	.16	*.61*	.27	.43	.18
7. Frequent improvements in fringe benefits—important	.26	*.60*	−.12	−.27	.09
8. Frequent improvements in fringe benefits—more	.34	*.53*	.37	.31	.11
9. Consideration and understanding from my boss—important	−.24	.05	.01	.16	.27
10. Consideration and understanding from my boss—more	.14	−.01	*.79*	.13	.16
11. Mutual trust with my boss—important	.04	.08	.07	.10	.40
12. Mutual trust with my boss—more	.14	.03	*.75*	.20	.09
13. Being given recognition for my efforts when deserved—important	.22	−.02	.28	−.03	−.08
14. Being given recognition for my efforts when deserved—more	.16	−.06	*.68*	.44	−.03
15. Being able to use my own approach on the job—important	−.10	.15	.33	−.12	.14
16. Begin able to use my own approach on the job—more	.03	.16	*.66*	.34	.23
17. Cooperative relations with my co-workers—important	.15	.36	−.34	.30	.31
18. Cooperative relations with my co-workers—more	.02	−.08	.21	*.71*	.16
19. Respect from my co-workers—important	.15	−.08	−.24	.22	.02
20. Respect from my co-workers—more	.07	.00	.29	*.65*	−.01
21. Making full use of my abilities at work—important	−.05	.21	.11	−.20	*.61*
22. Making full use of my abilities at work—more	.01	.06	.48	.20	*.60*
23. Developing new skills and knowledge at work—important	.21	−.02	−.05	.01	*.76*
24. Developing new skills and knowledge at work—more	.27	.06	.12	.21	*.66*
25. Being challenged by my work—important	.04	.05	.02	.02	*.71*
26. Being challenged by my work—more	.00	.04	.24	.18	*.72*
27. Opportunities to invent and explore new ways of doing the job—important	.17	.30	.01	.32	.30
28. Opportunities to invent and explore new ways of doing the job—more	.14	.77	.33	.55	.27

Note.—The highest loadings on each factor which are used to define the factor are italicized.

aspect. Items 27 and 28, dealing with inventing and exploring on the job, did not load highly on any of the factors. They had been expected to be part of the Growth factor. Table 6 gives the median correlation and number of items for each desire variable measured by rating scale. Table 7 presents the correlations between the two interview coders for each of the desire measures.

Table 8 gives the multitrait-multimethod matrix for rating-scale and interview desire measures. The median rating-scale reliability coefficient is .77; the comparable figure for interviews is .76. Two of the rating-scale reliabilities—Fringe Benefits and Respect by Peers—are quite low. With the relatively low reliabilities one would not expect to find high coefficients on the validity diagonals, and, in fact, this is what the data show. None of the entries on the validity diagonal is very high;

TABLE 6

RATING-SCALE DESIRE MEASURES: NUMBER OF ITEMS AND MEDIAN CORRELATION

Factor	No. items	r
Pay	4	.45
Fringe Benefits	4	.31
Respect by Superiors	4	.56
Respect by Peers	2	.52
Growth	6	.43

Note.—N = 176.

TABLE 7

INTERVIEW DESIRE SCALES: INTERCODER CORRELATION

Scale	r
Pay	.79
Fringe Benefits	.61
Respect by Superiors	.66
Respect by Peers	.61
Use Skills and Abilities	.51
Learn Skills and Abilities	.54

Note.—$N = 75$.

there is little evidence of convergent validity. Only the Respect by Superiors coefficient meets the first criterion for discriminant validity; it is higher than any other entry in its row and column. It also meets the second criterion for discriminant validity in that it is higher than any correlation between it and other traits measured either by interview or by questionnaire. In general, Desire for Respect by Superiors is uncorrelated with other traits. It is, thus, the only trait for which discriminant validity can be established by rating-scale and interview methods.

A fully counterbalanced paired-comparison method was also used to measure desires.[5] The first four variables were the same as in the preceding analysis. However, prior to the analysis the desire to use skills and abilities and the desire to learn skills and abilities were expected to be separate scales. They are reported separately in Table 9. However, in order to be consistent with the preceding analysis, a desire for growth scale was developed by adding together these two subscales; their correlation is .53. They appear as a single scale in Table 10.

Table 9 presents the split-half correlations for each of the paired-comparison desire measures. When the correlations are transformed into reliability estimates, the median

[5] A seventh scale was included in the design to reduce interscale bias. This scale was job security. In addition, the format of the paired comparisons did not require an S to choose which item was more important than the other. He also had the option of selecting a third alternative, neither one was more important than the other. In summing the choices for each alternative, one-half was given for each item when S indicated that he was indifferent between the two.

TABLE 8

DESIRE-RATING-SCALE MULTITRAIT-MULTIMETHOD MATRIX

	Questionnaire (n = 176)					Interview				
	P	FB	RS	RP	G	P	FB	RS	RP	G
Questionnaire										
P	(.77)									
FB	.00	(.47)								
RS	.00	.00	(.84)							
RP	.00	.00	.00	(.68)						
G	.00	.00	.00	.00	(.82)					
	(n = 58)					(n = 75)				
Interview										
P	.24	.17	.04	.07	.03	(.88)				
FB	.06	.03	.01	−.04	−.25	.09	(.76)			
RS	.00	.02	.22	−.13	−.10	.14	−.02	(.80)		
RP	−.02	.28	−.10	−.16	−.02	.10	−.02	.12	(.76)	
G	.28	.04	−.10	.09	.02	.04	−.03	.08	.03	(.60)

Note.—See Note of Table 4.

TABLE 9

SPLIT-HALF CORRELATIONS FOR PAIRED-COMPARISON IMPORTANCE MEASURES

Measure	r
Pay	.87
Fringe Benefits	.88
Respect by Superiors	.89
Respect by Peers	.86
Use Skills and Abilities	.84
Learn Skills and Abilities	.80

Note.—N = 138.

value is .93. This is considerably higher than the median reliability of the comparable rating-scale measures.

Table 10 gives the multitrait-multimethod matrix for the paired-comparison questionnaire and interview. Like the rating-scale coefficients, none of the values is very high, although their median value is .22, compared to .03 for the rating-scale procedure. With .35 and .37 coefficients, respecitvely, one could suggest that there is limited evidence for convergent validity in benefits and growth scales.

The Pay, Benefits, Peers, and Growth traits are higher than any entry in their respective rows and columns where neither trait nor method is common. These four scales, therefore, meet the first criterion of discriminant validity. Additional evidence for the discriminant validity of these scales is partially blocked, however, by the property of the paired-comparison scales that when an S chooses one item in a pair he automatically chooses against another item. Correlations among the scales are thus developed artifactually. There is an identifiable pattern among these correlations found in the questionnaire-questionnaire section of Table 10. Pay and Fringe Benefits correlate positively with each other but negatively with Respect by Superiors, Respect by Peers, and Growth. The Respect by Superiors and Respect by Peers scales correlate positively with each other and have essentially zero relationship with the Growth scale. Off-diagonal entries in the other portion of the matrix tend to approach zero.

DISCUSSION

The results of this study have implications for two general issues. One deals with the measurement problems associated with the traits under consideration in the investigation. Which variables have the greatest validity overall? What implications for research design does this have? The second point pertains to what the results say about the interview versus questionnaire controversy. To what extent can one say that one method is more effective than or equivalent to the other?

With the possible exception of the Respect by Peers trait, the reliability and convergent and discriminant validity of the satisfaction scales are quite high. Respect by Superiors has the highest convergent validity coefficient and it is quite closely followed by Growth. When there tends to be zero correlation among variables, as there was in this study, a high convergent validity coefficient is tantamount to high discriminant validity. But the same uncorrelatedness among variables makes

TABLE 10

DESIRE PAIRED-COMPARISON SUB MULTITRAIT-MULTIMETHOD MATRIX

	Questionnaire (n = 138)				
	P	FB	RS	RP	G
Questionnaire					
P	(.93)				
FB	.32	(.94)			
RS	−.48	−.44	(.94)		
RP	−.42	−.33	.43	(.93)	
G	−.44	−.50	.07	.06	(.90)
	(n = 50)				
Interview					
P	.21	.16	−.18	−.14	−.10
FB	−.06	.35	−.04	.02	−.23
RS	.01	.32	−.12	−.10	.01
RP	−.17	.03	−.08	.22	−.06
G	.07	−.14	.07	−.31	.37

Note.—See Note of Table 4. The Interview-Interview entries would be the same as in Table 8, so they are not repeated here.

it possible to attain discriminant validity when convergent validity is not high. This is what happened for the Satisfaction of Respect by Peers trait. A possible explanation for the low convergent validity of Respect by Peers satisfaction is suggested by the fact that some managers, when asked about their co-workers, spoke about their subordinates. It was possible to rephrase an interview question to obtain the desired data, but this, of course, could not be done for their filling out the questionnaire privately.

The reliabilities and convergent and discriminant validities of the desire scales were generally lower than for the satisfaction scales. Indeed, only the Fringe Benefits and Growth scales attained moderate convergent validity, and that was only when the paired-comparison method was used. By the rating-scale method, Desire for Respect by Superiors met the criteria for discriminant validity; none of the other scales did. Pay, Benefits, Peers, and Growth achieved some measure of discriminant validity by the paired-comparison method. On the surface, therefore, it would appear that the paired-comparison method is superior to rating scales for measuring desires—even though the statement is not strong because of the low convergent validity. In fact, however, this writer would recommend against using the paired-comparison method to measure desires, not because of its statistical properties, but because of the frustration and mistrust it causes among respondents. The rating-scale method was introduced into the study to find a less alienating method for obtaining data on desires.

Why the generally low validity of the desire scales? Are there ways this might be improved? Respondents frequently had trouble saying what was important to them when the question was asked directly in an interview. This was particularly true for the people holding lower-skill jobs. One could, however, learn quite a bit about what was important to these people by the subjects they brought up spontaneously. It was also true that more questionnaires had to be discarded for unanswered desire items than for omitted satisfaction items. This was more true for the paired-comparison items than for the rating-scale questions, but the statement applies to all desire measures.

Recall the pattern of factor loadings for the rating-scale desire factors in the interpersonal need areas. While both the "important" and "more" items loaded heavily on their respective factors for existence and growth needs, only the "more" items loaded on the Desire for Respect by Superiors and the Desire for Respect by Peers factors. For these interpersonal needs, people apparently do not consistently say the factor is important when they say more is desired. Some of the dynamics of this phenomenon are underlined by the behavior of one particular *S*. During his interview he described a rather lengthy series of difficulties which he had with his boss, both spontaneously and when specific questions were asked. His questionnaire reflected the lack of respect he felt from his superior. The desire items on his particular instrument were of the paired-comparison type, and not once did he select "understanding and trust with your boss" as being important to him. But he wrote the following note on the last page of his questionnaire:

> Supervision has to be improved in order for us to do a better job. This is in reference to all those questions pertaining to "understanding and trust from your boss."

The paradox of this event is that while his relationship to his boss was not "important" according to the way he filled out his questionnaire, it was important enough to deserve a special hand-written note on the instrument. Very few people wrote anything extra on the questionnaire. The point is to suggest that some measures—particularly those which ask directly about importance in interpersonal areas—appear to become less valid the more the person is experiencing difficulties in the area.

Since the "important "scales do not seem to work well to measure interpersonal desires, one could simply include a greater number of the "more" items for these areas. This is following the principle of test theory that one can increase the reliability of his test by making it longer. A similar point can be applied to the interviews, which were not very long anyway. It is not unusual for interviews to last 2 and 3 hr. in behavioral science re-

search. Longer interviews would allow for more issues to be raised spontaneously and for a greater sense of trust to develop between interviewer and respondent.

Perhaps the issues raised spontaneously by people in the interviews are not fully conscious desires. They may be more accurately termed preconscious desires. If this is true, it provides a partial explanation why the interview measure did not correlate highly with the more fully conscious questionnaire measures, while at the same time they did show evidence of discriminant validity. One might consider the possibility of using projective techniques to tap the desire variables; they could be used in both interview and questionnaire frameworks.

A subject not touched by this report is construct validity. How well do the different measurements confirm theoretically derived predictions? The instruments were developed to test predictions from a theory. In data already available, results show some, though not all, of the predictions are supported (Alderfer, 1966). Questionnaire variables relate to other questionnaire traits; interview measures to other interview scales; and questionnaire measures to interview variables. Replication of those findings is being undertaken. But even as the data stand now, they strongly suggest that questionnaire methods can be used to replicate interview results, and vice versa.

Particularly in approaching the interview versus questionnaire controversy, one would want to ask if the results of the study can be accounted for by design or demand characteristic artifacts. Each interview was carried out before the questionnaire was completed. There were anywhere from a few days to a few weeks lag between when a respondent had his interview and when he returned his questionnaire by mail to the writer. Since the interview questions were open-ended and the questionnaire items fixed-alternative, a respondent would have to have a remarkable memory and knowledge of the concepts of the study in order to produce correlated responses on the two instruments simply to appear consistent in his own and/or the eyes of the investigator.

The questionnaires were analyzed prior to the interviews, and both analyses were done blind so that neither questionnaire nor interview coder knew what the *S*'s other instrument scores had been. Furthermore, factor scores were used in most of the correlations; it would be quite an effort and ability to develop factors which are as clearly interpretable as those in this study through conscious or unconscious bias.

The results do more to answer the objections of those who say that interviews cannot be done systematically than they do to meet the points raised by those who say that questionnaires alienate respondents and thus foster invalid data. The respondents had met from a minimum of 20 min. to a maximum of an hour and a half with the researcher prior to being asked to do the questionnaire. They, thus, had had an opportunity to express their views in their own way and had less reason to feel exploited than people just asked to fill out a questionnaire.

Because it allows for more choice and initiative on the part of the respondent—without apparent loss of validity—the interview has much to recommend its being used in conjunction with questionnaires when the possibility following the study with organizational change activities exists. In time and dollar economics more data can be obtained per unit of output by using questionnaires. When interviews can be used in conjunction with questionnaires, there exists the possibility of reducing respondent-researcher alienation and thus increasing the validity of the questionnaire data.

The interview versus questionnaire controversy recalls the "two disciplines of psychology" described by Cronbach (1957). The disciplines were not bipolar but two dimensional: qualitative–quantitative being one dimension and correlation–experiment being the other. When essentially orthogonal issues become polarized, one can infer that emotional as well as intellectual issues are at stake. A report such as this which deals with the intellectual aspect is a limited contribution to the conflict.

REFERENCES

ALDERFER, C. P. Differential importance of human needs as a function of satisfactions obtained in the

organization. Unpublished doctoral dissertation. Yale University, 1966.

CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, **56**, 81–105.

CRONBACH, L. J. The two disciplines of scientific psychology. *American Psychologist*, 1957, **12**, 671–683.

DUNNETTE, M. D. Personnel management. *Annual Review of Psychology*, 1962, **13**, 285–314.

GLASS, G. V., & TAYLOR, P. V. Factor analytic methodology. *Review of Educational Research*, 1966, **5**, 566–587.

HECHINGER, F. M. A survey of surveys. *New York Times*, September 11, 1966. Section 4, p. 7.

HYMAN, H. H., & SHEATSLEY, P. B. "The Authoritarian Personality,"—a methodological critique. In R. Christie & M. Jahoda (Eds.), *Studies in the scope and method of the authoritarian personality.* Glencoe, Ill.: Free Press, 1954. Pp. 50–122.

KAISER, H. F. Formulas for component scores. *Psychometrika*, 1962, **27**, 83–87.

SEARS, R. R. Comparison of interviews with questionnaires for measuring mothers' attitudes toward sex and aggression. *Journal of Personality and Social Psychology*, 1965, **1**, 37–44.

(Received December 14, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 521–528

SELF-ESTEEM OF INDUSTRIAL WORKERS

JOEL LEFKOWITZ[1]

Baruch School, City College of New York

The Bills Index of Adjustment and Values (IAV) was used to obtain measures of self-esteem for a sample of 155 factory workers. The workers' self-esteem was significantly higher than college students' used as a standardization group for the IAV. Self-esteem proved unrelated to a variety of job-related variables. Married workers demonstrated greater self-esteem than unmarried. Those with some college experience demonstrated greater self-esteem than those with only a high school education. Employees who had to change their jobs demonstrated significant changes in level of self-esteem following the job change. Evidence was presented which suggests the possibility that those employees who respond readily to survey-research inquiries are lower in personal and job adjustment than are those who choose not to respond.

Since the industrial revolution, the worker has been successively defined by industry as motivated primarily by concern for physical comfort, economic success, the most efficient use of his talents, and emotional well-being (Brown, 1958; Herzberg, 1966). The prevailing contemporary view is that man is characterized by an ultimate goal to fulfill inherent potentialities and to gain some measure of competence over his environment. This "self-actualization" motive has been applied extensively to industrial life. Research indicates that one's job is potentially the greatest source of satisfaction or frustration of this need (Friedmann & Havighurst, 1954; Gurin, Veroff, & Feld, 1960; Levenstein, 1962).

Rarely made explicit is the importance of the individual's apperception of the extent to which he is "self-actualizing," as reflected by his feelings of self-esteem. Self-esteem is that organized and evaluative aspect of self which develops as a result of holding oneself up to scrutiny as an object in one's own cognitive world. This study is based upon the assumption that one's feelings of self-esteem are determined largely by reinforcement stemming from one's performance of his job, and from interpersonal relations on the job.

Problem

Measures of self-esteem were obtained for a sample of industrial workers in order to ascertain the distribution of scores of this variable for such a sample. A standardized measuring instrument was used in order to enable intra- and intergroup comparisons.

The second major purpose of the investigation arises from what has been called "an environmental approach to mental health in industry" (French & Kahn, 1962). The relationships between several job and demographic variables and level of self-esteem were analyzed.

Ever since the pioneering work of Coch and French (1948), workers' reactions to change have been shown repeatedly to be based on perceived threat (Mann, 1962; Marrow, 1949; McMurry, 1946; Schachter, Willerman, Festinger, & Hyman, 1961; Selekman, 1962; Shils, 1963; Trumbo, 1961; Walker, 1950). If one accepts the assumption that a man's job constitutes an important developmental and confirmatory source of many aspects of his self-concept, then it seems likely that these feelings of threat derive, in part, from the perception of threat to the self.

It was predicted that those employees who are moved to a different job would manifest significant changes in level of self-esteem (measured without regard to direction of change).

Directional changes were not predicted because we are interested primarily in noting whether or not an association exists between job change and change in level of measured self-esteem regardless of the positive or negative nature of that change. In addition, while

[1] Also at BFS Psychological Associates, Inc.

TABLE 1

DEMOGRAPHIC CHARACTERISTICS OF THE SAMPLE

Characteristic	M	SD
Age (years)	47.5	6.95
Company seniority (years)	21.7	5.20
Base rate of pay (cents/hr)	312.08	28.75
Education (highest grade completed)	11.25	1.42

Note.—Preliminary statistical tests revealed no significant differences between men and women on all of the major variables involved in the study. Data from men and women were therefore combined. $N = 179$; 144 males, 35 females.

the predominant reaction to job change is the perception of threat and attendant self-devaluative feelings, some individuals may manifest a defensive increase in self-esteem scores. Bills (1953) and Thorne (1954) note that self-devaluative feelings may result in increases in measured self-esteem among those who are initially low in measured self-esteem.

METHOD

Research Site and Sample

The study was conducted within two manufacturing divisions of a midwestern plant each of which employed approximately 275 hourly workers engaged in the manufacture of a variety of automotive, aircraft, and missile products. Of the 549 sets of questionnaires mailed, 179 were returned.[2] The relatively low return of 33% was due to the lengthy and difficult nature of the forms, and because of the necessity of using a direct-mail procedure. Table 1 summarizes some demographic characteristics of the sample.

Measures

Self-esteem was measured by the Bills Index of Adjustment and Values (IAV) (Bills, undated manual; Bills, Vance, & McLean, 1951). A self-esteem score is obtained by summing the respondents' self-ratings to 49 personality traits which are representative of items which occur frequently in client-centered interviews. Published reviews indicate the IAV to be the most reliable and valid measure of self-concept variables (Strong & Feder, 1961; Wylie, 1961).

The odd-even reliability of the self-esteem ratings was .92 (with Spearman-Brown correction).

Job satisfaction was measured by summing the responses to 30 Likert-type questions with 5 alternatives. The individual items represent 10 areas of job satisfaction which have been identified as subsuming all of the research results in this area with a minimum of overlap (Herzberg, Mausner, Peterson, & Capwell, 1957).

The odd-even reliability of the job-satisfaction questionnaire was .89 (with Spearman-Brown correction). The mean item-total correlation for the 30 questions was .56 ($SD = .18$).

During the course of the investigation it was recognized that the job-satisfaction items were concerned predominantly with extrinsic job aspects. Therefore, a separate job-satisfaction score was computed for the five items which related to intrinsic job factors. Herzberg's (1966; Herzberg, Mausner, & Snyderman, 1959) conceptualization of "motivators" was utilized to define the intrinsic items.

Employees' attitudes toward work-related change were measured by means of a change scale taken from Trumbo (1961). The scale consists of nine questions designed to measure generalized attitudes toward work-related change.

The respondents' ego involvement in their jobs was measured by summing the responses to three questions. The first question is based on research reported by Slater (1960) and Vroom (1962) utilizing Zeigarnik's work regarding recall of incompleted tasks. The question asks, "If a problem comes up in your work and it isn't all settled by the time you go home, how likely is it that you will find yourself thinking about it after work?" The second question is similar to one used by Gurin et al. (1960). The *Ss* were asked "What things do you particularly like about your job?" The write-in responses were scored independently by three judges for degree of ego involvement. The third question required the respondent to rank five areas of his life in their order of importance to him. These were family, job, hobbies and leisure-time activities, friends, and religious life. The *Ss'* scores were determined by the rank assigned to the job.

Computation of both inter-item and item-total correlations indicated that the three questions were statistically independent, suggesting that "job involvement" is a multidimensional trait. This agrees with results reported by Lodahl and Kejner (1965) subsequent to the initiation of the present investigation.

Skill level was estimated by base rate of pay. Base rate excludes from consideration all bonuses such as allowances for working second or third shift.

Attendance was measured by summing the number of times an employee was absent during the previous calendar year. An absence of more than 1 day's duration was counted as 1. It was assumed that this measure would be a rough indicator of attitudinal factors leading to a temporary withdrawal from the work situation.

The total number of voluntary visits to the plant dispensary during the previous calendar year was noted. Kasl and French (1962) report significant relationships between the tendency to seek medical aid and to report physical symptoms with

[2] Twenty-four individuals neglected to fill out the Bills Index of Adjustment of Values, so that the self-esteem data are based upon an N of 155.

skill level, self-acceptance, and a self-report measure of stress.

Procedure

The questionnaires were mailed directly to the homes of the respondents. A return envelope addressed to the investigator at Western Reserve University was enclosed. Follow-up letters were sent to those employees who had not replied within 2 wk.

Because of the nature of operations at the factory where this research was carried out it was possible to readminister the questionnaires to 19 employees who were "bumped" to different jobs from the ones in which they were situated at the time of the first administration of the questionnaires.

Most contracts at this factory run for approximately 6–16 mo. with the possibility of extension. Having to change one's job is, therefore, not an infrequent event. These changes are accomplished through the company's plant-wide "bumping" system, which seeks to accommodate the entire plant's overall manpower requirements.

When a particular job is discontinued (usually due to termination of a contract), the occupant of that job may move to the same job in another department. If none is available, he moves to the next lowest-ranking job for which he is qualified, thereby "bumping" the man with the lowest plant-wide seniority on that job. This man, in turn, continues the process and many men may eventually be "bumped." This continues until, if there are no more displacements possible, the man with least seniority is laid off. A plant-wide "bump sheet" which lists all necessary moves is printed daily.

A control group was obtained by readministering the questionnaires to 42 employees who had remained on the same job throughout the period of investigation.

Statistical Treatment

The major portion of the data was treated by multiple-regression analysis accomplished through the use of MURA (Hammer, Hansen, & Haynam, 1964), a stepwise computer program which is designed to handle up to 61 variables.

RESULTS

Preliminary Checks for Possible Sample Bias

Because only 33% of those questionnaires mailed were returned, the possibility exists that systematic "self-selection" effects have biased the sample.

The difference in mean seniority between the sample (21.7 yr.) and the population from which it was drawn (21.2 yr.) was not significant.

There is, of course, no data with respect to the major variables involved in the study for those employees who did not respond to the questionnaires. However, of the 179 employees who did respond, 63 did so only after receiving a follow-up (reminder) letter. This group may be used to approximate a group of "nonrespondents." It is possible that those who were reluctant to answer questionnaires differ significantly on relevant variables from those who responded readily.

Table 2 shows that results for five of the six variables were not significant, suggesting the absence of self-selection bias. However, there is a tendency for the follow-up respondents to be on higher-paying (more skilled) jobs, and to demonstrate greater job and personal adjustment. The possibility therefore exists that this trend is more pronounced among those who were most reluctant to answer the questionnaires—that is, those who chose not to respond at all.

Correlates of Self-Esteem

In order to ascertain the degree of association of each of the independent variables

TABLE 2

COMPARISON OF "EARLY" RESPONDENTS WITH THOSE WHO RESPONDED ONLY AFTER RECEIPT OF A FOLLOW-UP LETTER

Variable	Early respondents ($N = 116$)	Follow-ups ($N = 63$)	t
Self-Esteem[a]			
$\bar{X}$	179.62	183.15	*ns*
SD	26.70	23.56	
Job Satisfaction			
$\bar{X}$	102.77	105.87	*ns*
SD	16.70	16.28	
Job Satisfaction "Motivators"			
$\bar{X}$	15.40	16.94	2.80*
SD	3.45	3.71	
Attitude toward Work-Related Change			
$\bar{X}$	29.37	30.16	*ns*
SD	6.22	5.74	
Ego Involvement in Job			
$\bar{X}$	32.27	33.59	*ns*
SD	5.32	6.01	
Base Rate of Pay (cents/hr)			
$\bar{X}$	309.84	316.21	*ns*
SD	28.80	28.67	

[a] Due to missing data the *N*s are 101 and 54, respectively, for this variable.
* $p < .01$ ($df = 177$).

TABLE 3

STEP-WISE MULTIPLE REGRESSION ON LEVEL OF SELF-ESTEEM

Step no.	Regression coefficients											Constant term	R
	1	2	3	4	5	6	7	8	9	10	11		
1	.83											167.68	.12
2	.82	.04										156.54	.16
3	.69	.04	.08									132.59	.18
4	.65	.04	.10	.89								127.61	.20
5	.67	.05	.08	.93	1.09							128.39	.21
6	.67	.05	.09	1.01	1.24	−.29						133.42	.22
7	.66	.04	.09	.94	1.32	−.25	−.91					142.59	.22
8	.65	.05	.09	.83	1.06	−.29	−.86	−.17				149.88	.23
9	.63	.05	.09	.83	1.12	−.29	−.83	−.16	.10			147.01	.23
10	.55	.05	.09	.82	1.16	−.29	−.86	−.16	.10	.03		145.36	.23
11	.55	.05	.09	.83	1.15	−.29	−.86	−.17	.10	.03	−.06	145.33	.23
β	.08	.13	.10	.07	.07	−.07	−.05	−.05	.02	.02	−.01		

Note.—$N = 155$.

with level of self-esteem, a step-wise multiple-regression analysis was performed (see Table 3). The order of addition of the independent variables was determined, at each step, by which of the remaining variables would contribute most in explaining the total variance in self-esteem scores. The β-weights corresponding to the regression coefficients when all 11 independent variables are entered into the equation are also presented. The order of addition of the independent variables was as follows:

1. Job-satisfaction "motivators."
2. Seniority.
3. Base rate of pay.
4. Absences.
5. Number of dependents.
6. Attitude toward work-related change.
7. Education level.
8. Age.
9. Ego involvement in job.
10. Overall job satisfaction.
11. Dispensary visits.

At no step in the analysis does R approach a significant value. It is apparent that none of the variables investigated, or any combination of them, accounted for a significant portion of the variance in self-esteem scores for this sample.

The absence of significant correlations is in contradiction to a body of research which supports the existence of a positive relationship between job status and self-esteem and/or mental health (French, 1963; Kasl & French, 1962; Kornhauser, 1965; Zander & Quinn, 1962).

Men and women did not differ significantly on any of the variables considered (self-esteem, job satisfaction, attitude toward work-related change, and ego involvement in one's job).

The mean self-esteem score for married employees was 181.99 ($SD = 25.91$, $N = 136$) and for single people 167.01 ($SD = 19.88$, $N = 12$). The difference is significant at better than the .10 level. Single people are greatly restricted in achieving the gratifications associated with a deep interpersonal relationship. In addition, Gurin et al. (1960) point out that the role requirements for solitary statuses are much less clearly delineated than those for married people, and that this ambiguity probably generates adjustment problems; we would add: adjustment problems which are reflected in lower levels of self-esteem.

There were no differences in level of self-esteem between those working the first shift ($N = 101$), second shift ($N = 45$), and third shift ($N = 9$). However, third-shift workers scored significantly lower than did first-shift workers on overall job satisfaction. This dissatisfaction probably stems from the social

and familial adjustments that are required of those who work at night. All those in the present sample who were working the third shift were married.

Table 4 shows the mean self-esteem score for those employees who have had some college experience to be significantly higher than that of employees with a high school education. An overall positive relationship between level of education and self-esteem is not implied, however, because grade school graduates also scored higher than high school graduates (although the difference is not statistically significant).

Change in Level of Self-Esteem following Job Change

Table 5 shows that those employees who underwent job changes demonstrated significantly greater changes in level of self-esteem (without regard to direction of change) than did the control group who underwent no job changes.

As expected, when the change scores are analyzed algebraically, positive and negative changes tend to cancel each other out resulting in nonsignificant mean differences.

A difficulty in interpretation arises because changes in self-esteem scores in the directions predicted and actually observed (i.e., lowered scores for those initially high in self-esteem and a defensive rise in scores for those initially low in self-esteem) constitute changes which one would predict upon readministration of any measuring instrument. That is, we expect a regression toward the population mean upon retest due to unreliability of the measure even for the control group.

TABLE 4

SELF-ESTEEM SCORES OF EMPLOYEES DIFFERING IN LEVEL OF FORMAL EDUCATION ATTAINED

Level of education	M	SD	N
Grade school (ninth grade and below)	182.93	19.24	14
High school (tenth–twelfth grade)	180.14[a]	26.71	133
College (enrollment for any length of time)	189.50[a]	14.08	8

[a] $t = 1.70$ ($p < .10$).

TABLE 5

CHANGE IN LEVEL OF SELF-ESTEEM FOLLOWING JOB CHANGE

Group	M_{Diff}	SD	t
Experimental ($N = 19$)	18.00	13.41	2.34*
Control (no job change) ($N = 42$)	12.95	11.24	

Note.—Change scores computed without regard to sign.
* $p < .05$.

However, the interpretive task is facilitated if these changes are significantly greater for the experimental group than for the control group. The relevant regression equations indicate that this is so. The regression of change in self-esteem following job change on initial level of self-esteem is given by the equation $Y_R = 120.84 - .661X$. The regression of change in self-esteem on initial level of self-esteem for the control group (no job change) is $Y_R = 40.02 - .206X$. The difference between the two slopes is statistically significant ($t = -1.83$, $p < .10$).

DISCUSSION

The theoretical position which holds that the individual's job is a significant determinant of self-concept variables has been rather devoid of "hard" data substantiating that position. Self-concept theory holds that almost any interpersonal experience can result in the incorporation of some aspect of life into the self-hood. Particularly relevant, it is felt, is one's work experience (Allport, 1937).

It appears that the weakest part of self-concept theory is with regard to explaining how the various relevant aspects of one's psychological environment get incorporated into the self. Rogers (1951) is the most specific on this matter:

> Whether or not an object or an experience is regarded as part of the self depends to a considerable extent upon whether it is perceived as within the control of the self. Those elements which we control are regarded as a part of the self, but when even such an object as a part of our body is out of control, it is experienced as being less a part of the self [p. 497].

In terms of self-concept theory, job change is, therefore, perceived as threatening because of its signifying loss of control over a vital aspect of one's life.

The present finding, indicating significant changes in level of self-esteem following job changes, demonstrates the functional dependency of self-esteem on vocational variables. While the small sample size involved does serve to limit possible generalizations, considerations exist which lend added significance to the findings.

Having to change one's job is a relatively frequent and expected occurrence at the factory in which this research was conducted. We would expect the differential effects on similar experimental and control groups to be more clearly delineated within a population where change is not such a common and expected occurrence.

Second, the measure of self-esteem employed in the study is a standardized personality instrument, developed within the theoretical framework of nondirective counseling. The important fact is that the adjectives which comprise the IAV are not at all job-related. For job change to have resulted in significant changes on such a measure indicates the potency of circumscribed job variables in affecting basic personality characteristics.

The similarity of the regression equations presented emphasizes the necessary utilization of a control group in longitudinal research. The fact that the same directional changes are noted in both the experimental and control groups leads us to ascribe a portion of those changes as due to a regression of scores toward the population mean. However, the fact that the difference between the slopes of the two regression functions is statistically significant indicates that the changes in level of self-esteem following job changes cannot be accounted for totally on this basis. It appears that the statistical regression effect and the psychological effects are additive.

The mean self-esteem score for the sample of 155 industrial workers was 180.88 ($SD = 25.63$). This is significantly higher ($p < .01$) than the mean score of 1,728 college students used as a standardization group for the IAV (Bills, undated manual).

The higher self-esteem scores of the industrial sample are probably related to the age differential between these groups, rather than to the difference in educational level. The industrial sample is a middle-aged group whose members have passed the major social and vocational crises of life. As one matures into middle age, major psychological adjustments are achieved, especially in terms of increasing congruity between aspirations and attainments. This increase in congruity results in greater self-esteem (acceptance of self).

Administration of the IAV to other vocational groups would be necessary to determine whether our factory sample is relatively high or low in self-esteem when compared with other vocational categories. The existing evidence (Kornhauser, 1965) suggests that factory workers are lower in self-esteem than comparable groups (utilizing a more subjective, qualitative index than the IAV).

The failure to note significant relationships between self-esteem and the job-related variables included in the regression analysis may be attributed to the homogeneity of the sample. Table 1 reveals a lack of dispersion with respect to pay and educational level, as reflected in the small sample standard deviations of these distributions. Moreover, the relatively large standard deviations with respect to age and seniority are misleading because these dispersions are not normally distributed. Only 20 employees of the total 179 are below 40 yr. of age, and none is below 30. The lack of dispersion serves to attenuate any relevant correlation coefficient.

The restriction in range is not due to inappropriate sampling, but reflects accurately the composition of employees at the factory in which the research was conducted. Manpower reductions over the past 10 yr. have resulted in a reduction in the total number of hourly employees from approximately 7,000 in 1956 to 2,600 at the initiation of the study. This has been at the expense of those younger workers with the least seniority.

In addition, we have reason to suspect that the dispersions of obtained scores on the psychological measures employed are also at-

tenuated. In checking for possible sample bias, evidence was noted (Table 2) which indicates the possibility that the 67% of employees who did not respond to the questionnaires represent those who might score at the upper ranges on the measures used.

The possibility of such self-selection bias has significant implications for the practice of industrial survey research in general. It suggests that nonrespondents may be psychologically healthier, better adjusted to their jobs, more ego involved in their jobs, better able to adapt to changing work situations, and more highly skilled than those who respond voluntarily to inquiries regarding industrial psychological research. One possible interpretation is that such research serves as a type of "escape valve" for the reduction of tensions and anxieties and as a means of expression of dissatisfaction with one's job for those workers so characterized. The better adjusted, more satisfied workers may respond only to the extent that they are willing to "cooperate in the interests of science."

Industrial psychologists may be developing a discipline based, in large measure, upon atypical samples.

REFERENCES

ALLPORT, G. W. *Personality: A psychological interpretation.* New York: Henry Holt, 1937.

BILLS, R. E. Manual for the index of adjustment and values, adult and high school senior form. University of Alabama, undated.

BILLS, R. E. A validation of changes in scores on the index of adjustment and values as measures of changes in emotionality. *Journal of Consulting Psychology,* 1953, 17, 135–138.

BILLS, R. E., VANCE, E. L., & MCLEAN, O. S. An index of adjustment and values. *Journal of Consulting Psychology,* 1951, 15, 257–261.

BROWN, J. A. C. *The social psychology of industry.* Baltimore: Penguin Books, 1958.

COCH, L., & FRENCH, J. R. P., JR. Overcoming resistance to change. *Human Relations,* 1948, 1, 512–532.

FRENCH, J. R. P., JR. The social environment and mental health. *Journal of Social Issues,* 1963, 19(4), 39–56.

FRENCH, J. R. P., JR., & KAHN, R. L. A programmatic approach to studying the industrial environment and mental health. *Journal of Social Issues,* 1962, 18(3), 1–47.

FRIEDMANN, E. A., & HAVIGHURST, R. J. *The meaning of work and retirement.* Chicago: University of Chicago Press, 1954.

GURIN, G., VEROFF, J., & FELD, S. *Americans view their mental health.* New York: Basic Books, 1960.

HAMMER, R. E., HANSEN, G. J., & HAYNAM, G. E. *MURA (multiple regression analysis).* Cleveland: Case Institute of Technology, 1964.

HERZBERG, F. *Work and the nature of man.* Cleveland: World Publishing, 1966.

HERZBERG, F., MAUSNER, B., PETERSON, R. O., & CAPWELL, D. F. *Job attitudes: Review of research and opinion.* Pittsburgh: Psychological Service of Pittsburgh, 1957.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. B. *The motivation to work.* New York: Wiley, 1959.

KASL, S. V., & FRENCH, J. R. P., JR. The effects of occupational status on physical and mental health. *Journal of Social Issues,* 1962, 18(3), 67–89.

KORNHAUSER, A. *Mental health of the industrial worker.* New York: Wiley, 1965.

LEVENSTEIN, A. *Why people work: Changing incentives in a troubled world.* New York: Crowell-Collier, 1962.

LODAHL, T. M., & KEJNER, M. The definition and measurement of job involvement. *Journal of Applied Psychology,* 1965, 49, 24–33.

MANN, F. C. Psychological and organizational impacts of technological change. Paper read at American Assembly Roundtable in Automation, Columbia University, New York, May 3–6, 1962.

MARROW, A. Human factors in production. *Personnel,* 1949, 25, 341–349.

MCMURRY, R. N. The problem of resistance to change in industry. *Journal of Applied Psychology,* 1946, 30, 525–564.

ROGERS, C. R. *Client centered therapy: Its current practice, implications and theory.* New York: Houghton Mifflin, 1951.

SCHACHTER, S., WILLERMAN, B., FESTINGER, L., & HYMAN, R. Emotional disruption and industrial productivity. *Journal of Applied Psychology,* 1961, 45, 201–213.

SELEKMAN, S. M. Shop changes and the security of the individual. In C. R. Walker (Ed.), *Modern technology and civilization.* New York: McGraw-Hill, 1962. Pp. 308–315.

SHILS, E. B. *Automation and industrial relations.* New York: Holt, Rinehart & Winston, 1963.

SLATER, C. W. R. Some factors associated with internalization of motivation toward occupational role performance. *Dissertation Abstracts,* 1960, 21, 373.

STRONG, D. J., & FEDER, D. D. Measurement of the self-concept: A critique of the literature. *Journal of Counseling Psychology,* 1961, 8, 170–178.

THORNE, R. B. The effects of experimentally induced failure on self-evaluations. *Dissertation Abstracts,* 1954, 14, 1817.

TRUMBO, D. A. Individual and group correlates of attitudes toward work-related change. *Journal of Applied Psychology,* 1961, 45, 338–344.

VROOM, V. H. Ego involvement, job satisfaction, and job performance. *Personnel Psychology,* 1962, 15, 159–177.

WALKER, C. R. *Steeltown: An industrial case history of the conflict between progress and security.* New York: Harper, 1950.

WYLIE, R. C. *The self-concept: A critical survey of pertinent research literature.* Lincoln: University of Nebraska Press, 1961.

ZANDER, A., & QUINN, R. The social environment and mental health: A review of past research at the Institute for Social Research. *Journal of Social Issues,* 1962, 18(3), 48–66.

(Received December 15, 1966)

Manuscripts Accepted for Publication in the *Journal of Applied Psychology*

Selection of Salesmen by Means of an Assessment Center: Douglas W. Bray and Richard J. Campbell*: American Telephone and Telegraph Company, 195 Broadway, New York, New York 10007.

Faculty Satisfaction and the Dean's Influence: An Organizational Study of Twelve Liberal Arts Colleges: Jerald G. Bachman*: Institute for Social Research, Survey Research Center, The University of Michigan, Box 1248, Ann Arbor, Michigan.

Effects of Changes in Job-Satisfaction Levels on Employee Turnover: Charles L. Hulin*: Department of Psychology, University of Illinois, Urbana, Illinois 61801.

Achievement Values and Scientific Productivity: Bishwa Nath Mukherjee*: Department of Psychology, Nagpur University, Nagpur, India.

Comparison of Behavioral Styles between Entering and Graduating Students in Officer Candidate School: Peter B. Petersen and Gordon L. Lippitt*: 2131 G Street, Building Y, George Washington University, Washington, D. C. 20006.

Team Performance as a Function of Team Arrangement and Work Load: William A. Johnston* and George E. Briggs: The Human Performance Center, Ohio State University, 404-B West Seventeenth Avenue, Columbus, Ohio 43210.

Expectations and Dropouts in Schools of Nursing: Mildred Engberg Katzell*: Director, Evaluation Services, National League for Nursing, Inc., 10 Columbus Circle, New York, New York 10019.

Effect of Percentage and Specificity of Feedback on Choice Behavior in a Probabilistic Information-Processing Task: Irwin L. Goldstein, Joseph T. Emanuel, and William C. Howell*: The Ohio State University, Human Performance Center, 404-B West Seventeenth Avenue, Columbus, Ohio 43210.

Effect of Lunar Gravity on Man's Performance of Basic Maintenance Tasks: Richard J. Shavelson* and Joseph L. Seminara: Biotechnology, Org. 55–60, Building 151, Lockheed Missiles and Space Company, P. O. Box 504, Sunnyvale, California.

Supervisory Influence and Work-Group Performance: Kurt R. Student*: The University of Michigan Institute for Social Research, Box 1248, Ann Arbor, Michigan.

Influence of Training, Method, and Relationship on the Halo Effect: Eva Metzger Brown*: Apartment 134, 115–25 Metropolitan Avenue, Kew Gardens, New York 11418.

Reaction Time as a Function of the Cue Properties of an Auditory Display: J. Richard Simon*: Department of Psychology, The University of Iowa, Iowa City, Iowa 52240.

* Asterisk indicates author for whom the address is supplied.

Journal of Applied Psychology
1967, Vol. 51, No. 6, 529–532

DECISION MODELS FOR CREDIT SCORING SYSTEM CUTOFF DETERMINATION

JAMES H. MYERS [1]

School of Business Administration
University of Southern California

Numerical scoring systems for evaluating retail credit risk provide a more objective basis for screening credit applications than traditional methods. The business firm using a scoring system is faced with the problem of determining the cutoff score which will maximize profits per sale, taking into account the ratio of good to bad accounts without a scoring system and the profit-loss margins on each transaction. This paper presents a solution to this problem using statistical decision theory.

Several publications have discussed the development and use of numerical scoring systems for evaluating credit applications at the retail level (Myers, 1963; Myers & Cordner, 1957; Myers & Forgy, 1963; Smith, 1964). While research models and methodologies have varied somewhat, most investigators have approached the problem in a similar manner. The objective has been to produce scores or weights for various personal-history and financial-background items (e.g., age, time at present job, current installment indebtedness) which can be added together to produce a single "risk" score—the higher the score, the more likely the applicant possesses the personal and financial qualifications which predispose toward prompt payment of credit obligations.

The typical successful numerical scoring system produces results similar to those shown in Table 1. These results came from an actual study in a personal loan chain in Los Angeles (Myers & Cordner, 1957), using accounts which had been accepted on the basis of subjective evaluation by experienced credit interviewers. Table entries were produced by scoring a sample of equal numbers of good (paid-up) and bad (delinquent) accounts on each of 17 personal-history items. Weights used for each item were developed from a previous sample of accounts, so that results shown in Table 1 are "shrunken" estimates of the discrimination power of the scoring system.

[1] The writer wishes to thank Harry V. Roberts, University of Chicago, for help in the preparation of this manuscript.

The score scale shown has been simplified from the original scale, which ran from 0 to 130 points.

Since all of the accounts scored had already been granted credit on the basis of a subjective evaluation, it is clear that the scoring system offers some *improvement* over subjective evaluation; for example, by establishing a lower cutoff score of 5 and below, 28% of the accounts which did not pay could have been eliminated at a cost of only 5% of good accounts.

Problem

A major problem in using such a system is that of determining the cutoff score below which applicants should be refused credit. One approach would be to reject those scoring 3 or below, which would eliminate approximately 6% of unpaid accounts without losing any good business. But the optimum solution is not so simple. It can be shown that the most profitable lower score limit is much higher than this in some situations; that is, the merchant can afford to eliminate some potentially profitable (good) accounts when by doing so he can eliminate a *much larger percentage* of potentially unprofitable (delinquent) accounts.

In general, there are two factors which must be considered in setting the most profitable lower cutoff score for any credit situation:

1. Gross profit margins on transactions. Obviously, the businessman would do better to set the cutoff higher (i.e., refuse many more applicants) if his gross profit margin

TABLE 1

CREDIT SCORES OF GOOD AND BAD ACCOUNTS

Score	% of good accounts[a]	% of bad accounts[a]
14		
13	3	
12	8	
11	10	3
10	14	6
9	23	9
8	19	12
7	13	17
6	5	25
5	3	15
4	2	7
3		4
2		2
1		
0		
	100%	100%

[a] $N = 150$.

were small. In such a situation, he stands to lose a great deal (80–90% of sale price) on each transaction. Banks are examples of this; they risk $100 to gain $6 or $7, in the case of short-term personal loans.

2. Percentage of all credit transactions that pay in full (without using a scoring system). Referring to Table 1, it can be seen that *if* the ratio of all paid to all unpaid credit accounts is 1:1, the cutoff score can be set higher than if the ratio were 15:1 or 20:1, as in many credit operating situations. In the latter case, moving the cutoff up eliminates a larger *percentage* of unpaid accounts but at the same time eliminates a far larger *number* of good accounts, on which potential profit is lost.

Statistical decision theory provides a means of determining the optimum cutoff score for any credit situation involving separate transactions on each of which profit or loss is readily determinable (e.g., finance-company loan, furniture-store purchase). The basic principles of decision theory are developed in several texts (Chernoff & Moses, 1959; Schlaifer, 1959).

Decision Models

Incremental analysis will determine the point on the score scale where the loss from rejecting good accounts first exceeds the gain from rejecting delinquent accounts. For each possible score i on the score scale ($i = 1, 2, \cdots k$) we can assign a probability P_{ig} that a potential customer will have score i *and* will pay, and probability P_{ib} that he will score i *and* will not pay. Let

$$P_{.g} = \sum_{i=1}^{k} P_{ig}$$

$$P_{.b} = \sum_{i=1}^{k} P_{ib}$$

$$P_{i.} = P_{ig} + P_{ib}, \quad i = 1, 2, \cdots k$$

Note that:

$$P_{.g} + P_{.b} = P_{1.} + P_{2.} + \cdots P_{k.} = 1.$$

and

$$P_{ig} = (P_{.g})[P(i|g)]$$

where

P_{ig} = joint probability of the occurrence of a "good" applicant *and* a score of i.
$P_{.g}$ = probability that a loan will be repaid (i.e., proportion of all loans which are repaid, without using a scoring system cutoff). This is a "prior" probability.
$P(i|g)$ = proportion of good (paid) cases which score i (from Table 1).

To represent profit for good transactions and loss on unpaid accounts, let

m_g = gross margin on a transaction which pays in full.
m_b = net loss on an account which does not pay.

Then we will accept applicants with score i if and only if

$$P_{ig}m_g \geq P_{ib}m_b, \qquad [1]$$

that is, if

$$\frac{P_{ig}}{P_{ib}} \geq \frac{m_b}{m_g}.$$

An alternative, but equivalent, formulation of the above decision rule is to accept applicants with score i if and only if

$$P(g|i)m_g \geq P(b|i)m_b, \qquad [2]$$

where $P(g|i)$ and $P(b|i)$ are conditional probabilities of repayment, given that the applicant has a credit score of i. However, since

$$P(g|i) = \frac{P(i|g)P.g}{P_{ig} + P_{ib}} = \frac{P_{ig}}{P_{i.}}$$

results from inequality, [2] will always be proportional to those from inequality [1] and the same cutoff decisions will be produced in either case.

Expected Profit

Perhaps a more meaningful approach is to calculate an *expected profit* from each credit sale or loan for each possible cutoff point on the score scale. The following model produces the same cutoff decisions as those above, but has the advantage of talking in terms that are more familiar and meaningful to the businessman. It also tends to smooth out irregularities in score distributions which occur when either good or bad cases (or both) pile up at a particular score or score interval.

If we assume, for the moment, that no scoring system is used, we have

$$E_o(\pi) = m_g \sum_{i=1}^{k} P_{ig} - m_b \sum_{i=1}^{k} P_{ib} \qquad [3]$$
$$= m_g P_{.g} - m_b P_{.b},$$

where m_g, m_b, $P_{.g}$, *and* $P_{.b}$ are as previously defined, and $E_o(\pi)$ = expected gross profit on an individual loan of average size, with no scoring cutoff.

For illustration, assume that the following values represent the operating situation in a personal loan office for a \$100 personal loan:

$$m_g = \$\ 15$$
$$m_b = \$100$$
$$P_{.g} = \ .95$$
$$P_{.b} = \ .05$$

Then Equation [3] yields the following expected profit from a \$100 personal loan for the normal operating situation without using a scoring system (assuming *no* payments are made on a bad loan):

$$E_o(\pi) = (.95)\ (\$15) - (.05)\ (\$100)$$
$$= 14.25 - 5.00$$
$$= \$9.25.$$

If a numerical scoring system is installed and used, it will mean that some of the accounts previously accepted will now be refused. The problem is one of finding the cutoff point which will maximize profit. One approach is to calculate the expected profit at each cutoff score by a variation of Equation [3]. If a cutoff is used, there are two forms of "profit":

1. Gross margins on good business which is *accepted;*
2. Losses saved by *rejecting* bad business previously accepted when no system was used.

Similarly, losses are realized both from bad business which is *accepted* and from good business which is *rejected* when the system is used. Expected profit can be calculated by:

$$E_j(\pi) = m_g \sum_{i=j+1}^{k} P_{ig} + m_b \sum_{i=1}^{j} P_{ib} - m_g \sum_{i=1}^{j} P_{ig} - m_b \sum_{i=j+1}^{k} P_{ib},$$

where $E_j(\pi)$ is the expected profit by using a scoring system and rejecting *all* applicants scoring j or below.

As an illustration, the expected profit per \$100 loan if applicants scoring 7 or below are rejected (using Table 1 and assuming 5% delinquent accounts without a scoring system) is:

$$E_7(\pi) = .77\ (.95)\ \$15.00 + .70\ (.05)\ \$100.00$$
$$- .23\ (.95)\ \$15.00 - .30\ (.05)\ \$100.00$$
$$= \$10.97 + \$3.50 - \$3.28 - \$1.50$$
$$= \$9.69.$$

In a similar manner, the expected profit can be calculated for each score on the scale, as shown in Table 2. From this table, it can be seen that the expected profit increases from \$9.25 without a scoring system to \$11.70 by using the system shown in Table 1 and eliminating all applicants scoring 6 or less. Thus, the cutoff should be set at 6 or below to maximize expected profit. This is the same cutoff point as would be determined using Equations [1] and [2].

TABLE 2

EXPECTED PROFIT AT EACH REJECTION POINT

Reject those scoring	$E_j(\pi)$
≤13	−9.25
≤12	−8.39
≤11	−6.12
≤10	−3.56
≤ 9	−.18
≤ 8	5.48
≤ 7	9.69
≤ 6	11.70
≤ 5	10.62
≤ 4	9.98
≤ 3	9.85
≤ 2	9.45
≤ 1	9.25
No system	9.25

Discussion

The above analysis does not consider, of course, the costs of constructing, installing, or administering a numerical scoring system. These costs would normally be small in relation to the increase in expected profit for a large volume operation using a scoring system as effective as the one illustrated in this paper.

It should be noted that the greater collection efforts normally required for those with lower scores would change the entries in the payoff table to some extent; that is, the gain per successful loan would be reduced and the loss per unsuccessful loan would be increased. Also, the \$100 loss figure used in this paper for a \$100 loan is not likely to be accurate for the practical situation. Usually the borrower will make at least *some* payments, thus reducing the loss for the average loan below this amount. Entries in the payoff table can easily be changed to reflect the individual company situation.

The analysis presented here was based upon loans of \$100. For loans at or near this amount, a linear utility function can be assumed, against which the calculated expected profits can be evaluated. However, this does not mean that the utility function would be linear throughout the entire possible size range of loans. A small operator would be reluctant to lend \$10,000 or more to any one individual, regardless of credit score. Thus, the shape and extent of the utility curve must be taken into account for each lending situation.

The greatest problem in determining proper credit scoring cutoffs is likely to be caused by the *preselection* exercised by experienced credit evaluators. As pointed out earlier, only applications which had previously been *accepted* are shown in Table 1 and have been used for the models in this paper. Payment probabilities of those who were *rejected* are simply not known, although they are assumed to be low. Since this is the case, the models presented here are, strictly speaking, to be used only *after* the customary subjective screening process has been carried on by credit evaluators.

REFERENCES

CHERNOFF, H., & MOSES, L. E. *Elementary decision theory*. New York: Wiley, 1959.

MYERS, J. H. Predicting credit risk with a numerical scoring system. *Journal of Applied Psychology*, 1963, 47, 348–352.

MYERS, J. H., & CORDNER, W. C. Improve credit operation profits. *The Credit World*, February 1957, 12–13.

MYERS, J. H., & FORGY, E. W. The development of numerical credit evaluation systems. *Journal of the American Statistical Association*, 1963, **58**, 799–806.

SCHLAIFER, R. *Probability and statistics for business decisions*. New York: McGraw-Hill, 1959.

SMITH, P. F. Measuring risk on consumer installment credit. *Management Science*, 1964, 11, 327–340.

(Received December 21, 1966)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 533–538

RELEVANCE OF PERSONAL NEED SATISFACTION FOR OVERALL SATISFACTION AS A FUNCTION OF SELF-ESTEEM[1]

ABRAHAM K. KORMAN

New York University

The purpose of this research was to test the hypothesis that need satisfaction was related to overall satisfaction for high-self-esteem individuals, but not for low-self-esteem individuals. The hypothesis was supported in 2 correlational studies and 1 experimental study. Implications of the research for theories of satisfaction and possible interpretation in terms of childhood learning experiences were discussed.

An often-proposed theoretical explanation as to why people choose a given behavioral role is that such choices are determined by the characteristics which a person sees in himself (i.e., his wants for that type of role) and which roles he believes will satisfy these personal wants and desires (cf. Vroom, 1964). To the extent, then, that the chosen role provides reinforcement for these personal wants and desires, this type of framework would predict increased liking for and satisfaction with that role. However, the possibility that people may choose roles because of their general acceptability and "social" characteristics, and like them on the same basis, independently of one's own desires, is not typically included in this framework, even though such phenomena occur and, sometimes, at cross-purposes with individual wants (Korman, 1967).

Thus, there is research which suggests that the "personal" model may be somewhat of an oversimplification. Blumenfield (1965) hypothesized that for individuals who had recently reenlisted in the Navy there would be a high correlation between their own values and the values they expected to be supported in the Navy. His correlations were, however, negative in direction ($r = -.18$ with $N = 108$, $p < .10$; $r = -.33$ with $N = 157$, $p < .01$). Similarly, Schletzer (1966) found no relationship between scores on Strong Vocational Interest Blank scales and satisfaction with one's occupation for six separate occupational groups.

In previous work (Korman, 1966, 1967) the writer has suggested and verified in several independent occupational choice situations the hypothesis that the "personal" model holds only for those high in self-esteem, which is defined as one's general evaluation of himself as a need-satisfying adequate individual. This was derived from the theoretical assumptions that individuals will find satisfying those situations which maximize their sense of cognitive balance and will engage in behavior appropriate to the attainment of those situations. One further theoretical assumption, not made explicit previously, is that there exists a social norm that individuals should find situations where their self-perceived needs are satisfied and where they perform well as more satisfying than situations where this is not the case.

The purpose of this paper is to report three further studies designed to extend this hypothesis to situations where the dependent variable in a situation is overall liking or satisfaction with it. Basically, it was hypothesized in all cases that for individuals of high self-esteem, situations where self-perceived needs are being satisfied are judged as being more satisfying than situations where one's self-perceived needs are not being satisfied. On the other hand, stemming from both theoretical assumptions above, it was hypothesized for low-self-esteem individuals that there is no relationship between satisfaction with a situation and the extent to which one's self-perceived needs are being met in that situation.

[1] This research was partially supported by a grant from the New York University Arts and Sciences Research Fund. Part of the data reported in this paper was presented at a meeting of the Eastern Psychological Association, Boston, April 1967.

STUDY ONE

The first study consisted of testing the above hypothesis in a study of social groups on the campus of a large private Eastern university. It was predicted on the basis of the above reasoning that an individual's overall satisfaction with a student group could be predicted by a knowledge of the discrepancy between the kinds of personal desires a person wanted satisfied in his group activities and the kinds of reinforcements he actually receives, but only for high-self-esteem individuals. Such predictive ability would, on the other hand, break down for low-self-esteem people.

Method

Sample. The sample consisted of 70 male and 57 female undergraduates at a large private Eastern university, a total of 127 in all. Since previous research by the author (Korman, 1967) has indicated that the theoretical approach seemed to be applicable to both sexes, the groups were combined for analysis.

Instruments. Each *S* was administered a set of instruments designed to test the hypothesis. These were as follows:

(*a*) *Self-Esteem*—Self-esteem was measured by the "Self-Assurance" scale of the Ghiselli (undated) Self-Description Inventory. This scale is designed to measure the extent to which the person sees himself as an adequate, competent, need-satisfying individual and has been used by the author in his previous investigations. Evidence for the construct validity of this scale is available in a number of papers (cf. Ghiselli, 1963, undated).

(*b*) *Social Needs*—Each *S*'s desires in a group situation and from a group situation were measured by a questionnaire entitled "The Ideal Group." This questionnaire consists of a list of a number of characteristics which might be typical of a group to which a college student might belong and each respondent is asked to rate on a 5-point scale how much each characteristic would be part of his/her "ideal group." Among these characteristics are "is highly oriented toward school and school activities," "is interested in religious activities," "is concerned about other people and is not self-centered," and "is interested in social events." The test-retest item reliability has a median of .80 (about 3 wk. apart).

(*c*) *Group Characteristics*—The characteristics which each individual perceived in his favorite group, that is, the one he spent most of his time with, were measured by a questionnaire entitled, "My Favorite Group." The items of this questionnaire were the same as for "My Ideal Group," but in a different order.

(*d*) *Group Satisfaction*—Each individual's overall satisfaction with his favorite social group, that is, the one he spent most of his time with, was measured by a 5-point Likert scale administered as part of a Personal Characteristics Blank which also indicated whether a person belonged to a group at all. The five steps of this scale, in descending order, were labeled as follows: (1) I find it extremely satisfying and interesting; (2) In general, I find it pretty worthwhile; (3) It's at least as good as most of the other groups I could spend my time with; (4) I would probably join a different group if I had the time to look around; (5) I don't know why I waste my time with them. These responses were scored on a 5-point basis from "most" to "least" satisfaction.

Procedure. (*a*) The questionnaires were administered during regular class meetings of Introductory Psychology classes.[2] In almost all cases, the "Ideal Group" and "My Favorite Group" questionnaires were administered in separate meetings averaging about 2 wk. apart. The other questionnaires were randomized as to presentation.

(*b*) Individuals were then grouped into "high" and "low" self-esteem categories, using the median for the national norms as the point of division.[3] Within each "Self-Esteem" category, the respondents were then grouped into "high" and "low" satisfaction categories on the basis of their responses to the 5-point Likert scale. The "high" group was defined to be those checking the highest scale category, while the "low" group was considered to be those checking Satisfaction Categories 1–4. Thus, the four separate cells which were developed and their respective *N*s were High Self-Esteem, High-Satisfaction ($N = 23$), High Self-Esteem, Low-Satisfaction ($N = 15$), Low Self-Esteem, High-Satisfaction ($N = 47$), and Low Self-Esteem, Low-Satisfaction ($N = 42$). While the High-Satisfaction groups were automatically equated across self-esteem by the grouping procedure used, two cases had to be eliminated from the High-Self-Esteem, Low-Satisfaction group in order to equate the two Low-Satisfaction groups on the satisfaction variable. As a result of withdrawing these two cases, the mean satisfaction score for the High-Self-Esteem, Low-Satisfaction group was 3.46 while the comparable mean for the Low-Self-Esteem, Low-Satisfaction people was 3.43.

(*c*) Average item-discrepancy scores were then computed between "desired" and "obtained" need satisfaction for each *S* utilizing the three highest-rated needs for each (four in the case of ties). These were computed by subtracting the obtained satisfaction from the desired.

---

[2] The author's appreciation goes to James Kirkpatrick and Donald Davis of New York University for allowing the use of their classes in obtaining *S*s.

[3] In previous studies, the point of division was 1/3–2/3, approximately. Since this approximated the median on the national norms, it was decided to adopt the more specific cutting point in these later studies. Although for some unexplained reason the average level of this group was somewhat below the national median, sufficient numbers of "highs" and "lows" did appear for analysis.

Results

The results of the investigation are summarized in Table 1 and indicate that both hypotheses were strongly supported. The average discrepancy between "desired" and "obtained" satisfaction shows a strong negative relationship to overall satisfaction, but only for the High-Self-Esteem group. On the other hand, there is almost no difference at all in average item-discrepancy scores for the Low-Self-Esteem groups, despite the fact that they claim differences in satisfaction with their groups. In fact, the two means are actually in reverse order from what would be predicted, although far from significantly so.

TABLE 1

SUMMARY OF RESULTS: AVERAGE ITEM DISCREPANCIES BETWEEN "DESIRED" AND "OBTAINED" CHARACTERISTICS FOR THOSE DIFFERING IN GENERAL SATISFACTION WITH THEIR SOCIAL GROUP AS A FUNCTION OF SELF-ESTEEM

General satisfaction	High self-esteem	Low self-esteem
High		
M	.469	.679
SD	.289	.485
N	23	47
Low		
M	.898	.642
SD	.723	.577
N	15	42
t	2.52*	.32
df	36	87
r_{bis}	−.54	.04

* $p < .01$.

STUDY TWO

The purpose of this study was to replicate the previous findings in the different, unrelated context of job satisfaction, a context which is perhaps of somewhat greater significance to the individual than that of satisfaction with a social group and where much previous theorizing has concentrated on the need-satisfaction model detailed in the introduction to this paper (cf. Dawis, England, & Lofquist, 1964; Schaffer, 1953). The prediction was similar to that in the previous study in that it was hypothesized that overall job satisfaction could be predicted on the basis of need-satisfaction discrepancies for high-self-esteem people, but not for those with low self-esteem.

Method

Sample. The sample consisted of 52 night-school students at a private Eastern university in a large city, all of whom were employed full time during the day and had been on the same job at least 6 mo. Of this sample, 29 were men but, as in the previous study, the combined group was used in the analysis.

Instruments. (*a*) Self-esteem was measured in the same manner as in the previous study, that is, by the Self-Assurance scale of the Ghiselli Self-Description Inventory.

(*b*) Individual job "desires" and what one perceived as being received from one's job were determined by administering the Crites Vocation Survey twice, once under a "normal" self-description set and once when they were asked to describe their jobs. This is a Likert-type instrument which is derived from factor-analytic research in this area and which has the following need scales: Material Security, Job Freedom, Structure, System, Personal Status, Behavior Control, and Social Service (Crites, 1961). The administration of the instruments took place 2 wk. apart.[4]

(*c*) "Overall" job satisfaction was measured by a specially developed scale of job satisfaction designed to measure the individual's overall "affect" toward his job. The scale consisted of five Likert-type items aimed at measuring the individual's general feeling toward his job. Since preliminary analysis indicated that these items showed the expected high intercorrelations the responses were summed for analysis purposes, thus providing a range of scores on "general job satisfaction" from 0 to 25.

Procedure. (*a*) The questionnaires were administered in two separate sessions 2 wk. apart. At the first administration the *S*s were asked to take home and fill out the Ghiselli Self-Description Inventory, the Crites Vocation Reaction Survey (self-set), and the Job Satisfaction Questionnaire. When they returned these 2 wk. later they were then asked to complete the Crites questionnaire under a job-description set. All *S*s were completing a research-participation course requirement.

(*b*) "High" and "low" self-esteem groups were derived as previously based on responses to the Self-Assurance scale on the Ghiselli Self-Description Inventory.

(*c*) "High" and "low" job-satisfaction groups were then derived on the basis of the generalized measure of job satisfaction within each self-esteem group separately, as in Study One. The four cells which were developed are shown in Table 2.

[4] The author's appreciation goes to Lawrence Karlin of the Department of Psychology at New York University for his assistance in this administration.

TABLE 2

MEAN SATISFACTION SCORES FOR HIGH- AND LOW-SATISFACTION GROUPS

Group	Mean job satisfaction	N
High Self-Esteem, High Satisfaction	21.00	13
Low Self-Esteem, High Satisfaction	20.50	9
High Self-Esteem, Low Satisfaction	17.16	17
Low Self-Esteem, Low Satisfaction	17.00	13

(*d*) Average need-discrepancy scores were then computed between "desired" and "obtained" need satisfaction for the two most important needs (three in the case of ties) in the same way as previously.

Results

The results of the investigation are summarized in Table 3 and provide support for the general hypothesis of the investigation; that is, the discrepancy between "desired" and "obtained" satisfaction provides a good basis for the prediction of overall job satisfaction for the High-Self-Esteem group, but not for the Low. While the sample involved is a small one, the fact that it provides excellent replication for the previous study but in a different area suggests the generality of the hypothesis offered may be fairly wide.

STUDY THREE

While the results of the two previous investigations provide strong support for the hypothesis derived from the theoretical model proposed in the introduction to this paper they both suffer from a flaw in that in both cases the characteristics of the environment (i.e., the "social group" and the job) have been described by the *S* himself, thus leading to certain ambiguities of interpretation. For example, it may be that individuals will describe their jobs or groups differently in terms of similarity to their self-description as a function of their liking for it, and this tendency might differ as a function of self-esteem. In order to resolve the difficulty, two separate studies were undertaken whereby the characteristics of an "environment" were varied independently by the investigator, and *Ss* were asked to indicate their degree of "liking" for it. Study A consisted of a set of verbal descriptions of different kinds of universities and the kinds of characteristics which they would emphasize to their students, while Study B consisted of a set of verbal descriptions of different kinds of college groups. The kinds of behaviors chosen for these descriptions generally conformed to several of the "needs" proposed by Murray (1938) and utilized in the Edwards Personal Preference Schedule (EPPS; Edwards, 1954). It was predicted that the high-self-esteem individual would be more likely to rate one "environment" over another as a function of his relative need-strengths. Hence, there should be a positive correlation between relative need-strengths and comparative ratings of "environments" for the High-Self-Esteem group. However, such a correlation should not exist for the low-self-esteem individuals.

Method

Sample. The *Ss* for Study A consisted of 45 undergraduate and 15 extension students, 37 of which were male. The sample for Study B consisted of 71 undergraduates, of which 43 were male. There was an overlap of 9 students between the two samples.

Instruments. (*a*) Self-esteem was, as above, measured by the Self-Assurance scale of the Ghiselli Self-Description Inventory (SDI).

(*b*) As a measure of self-perceived needs, the EPPS was utilized. This is a forced-choice instrument which is ipsatively scored, thus providing a measure of the relative strength of various needs within the individual.

(*c*) Questionnaires describing different universities (or college groups) were developed as a measure of

TABLE 3

SUMMARY OF RESULTS: AVERAGE ITEM DISCREPANCIES BETWEEN "DESIRED" AND "OBTAINED" JOB CHARACTERISTICS FOR THOSE DIFFERING IN JOB SATISFACTION AS A FUNCTION OF SELF-ESTEEM

Job satisfaction	High self-esteem	Low self-esteem
High		
M	2.10	4.09
SD	2.55	2.62
N	13	9
Low		
M	3.92	4.40
SD	2.00	2.31
N	17	13
t	2.04*	.19
df	28	19

* $p < .05$.

TABLE 4

STUDY A: CORRELATIONS BETWEEN AFFILIATION-ORDER NEED DISCREPANCY AND COMPARATIVE RATINGS OF "AFFILIATION-ORIENTED" AND "ORDER-ORIENTED" UNIVERSITIES

	High self-esteem	Low self-esteem
N	19	41
r	.65*	−.12

* $p < .01$.

different "environments." The descriptions for universities involved such as "emphasizes hard work and accomplishment," "emphasizes innovation and change," "realizes the importance of social relations," etc. The descriptions for college groups were such as "leading other groups," "making friends," "intellectual activity," etc. In each case *S*s were asked to indicate degree of liking for each situation on a 5-point scale.

Procedure. (*a*) The *S*s were administered the EPPS and the Ghiselli SDI at one class session. Several weeks later they were administered the "environmental description" blanks. In the first case, *S*s were told the data collection was part of a departmental project while in the second case they were told it was for university purposes.

(*b*) After deriving high and low self-esteem as previously, separate correlations were computed between level of need discrepancy on the EPPS for two needs and tendency to rate the corresponding environment over the other. For Study A the two needs studied were need for affiliation and need for order while for Study B the two needs were change and affiliation, with these variables chosen because of their relevance to the contemporary college environment. The verbal descriptions utilized were deliberately constructed in order that different "need pairs" could be utilized in the two studies.

Results

The results of the two studies are summarized in Tables 4 and 5 and in all cases indicate strong support for the hypothesis. That is, the correlations are significant in both cases for the High-Self-Esteem group but in neither case for the Low-Self-Esteem. While the results of Study A are somewhat more impressive than of Study B, both studies are supportive of the hypothesis.

DISCUSSION

The results of this investigation indicate that the notion that satisfaction is "personally determined" in that situations are evaluated according to the extent to which they are satisfying one's desires may be somewhat oversimplified. That is, it seems apparent that the meaningfulness of the "satisfied" response in terms of its correlates and possible determinants does not seem to be the same for groups differing in self-esteem.

Several questions stem from this finding, it would appear. For example, what determines the "satisfied" response for the low-self-esteem person if self-perceived need satisfaction does not? One possible hypothesis stems from our suggestion that the low-self-esteem person may be more subject to the influence of others in choosing a vocational role, even though the choice is not perceived as "self-fulfilling," since "self-fulfilling" or need-satisfying states are not in "balance" for him. A similar process may operate in the "satisfaction" situation of concern here in that the low-self-esteem person may evaluate a situation in terms of how he perceives others rate it and what its general social "acceptability" is, rather than in terms of how the situation meets his needs. Thus, for example, a person of low self-esteem may see a situation as not meeting his needs at all and yet might say he was very satisfied with it, if he thought that other people in the situation were "happy" with it. On the other hand, he might rate it unsatisfactorily if he thought other people were unhappy with it, independently of how he himself felt about it. In this sense, then, one might argue that the low-self-esteem person is "externally oriented" in that he generally desires to go along with others independently of what he might want in a specific situation and across different situations, while the high-self-esteem individual is under "internal orientation" in that the world is evaluated according to his own desires and his own self-perceived needs in a specific situation.

TABLE 5

STUDY B: CORRELATIONS BETWEEN CHANGE-AFFILIATION NEED DISCREPANCY AND COMPARATIVE RATINGS OF "CHANGE-ORIENTED" AND "AFFILIATION-ORIENTED" COLLEGE GROUPS

	High self-esteem	Low self-esteem
N	29	42
r	.39*	.24

* $p < .05$.

A second question of considerable theoretical interest, but also of importance in a practical sense, is why should "consistency" be so important, since it seems to fly so much in the face of "common sense"? One possible interpretation may be suggested in the area of childhood training. This is that the individual is encouraged to be consistent and noncontradictory in his behavior by both social norms and parental strictures. He thus seeks out and prefers situations which reduce the anxiety which would develop as a result of violating such norms not to be "contradictory," even though such situations may actually be nonbeneficial to himself. This kind of behavior would also be predicted to be more likely the more one were to be susceptible to these norms, a characteristic commonly associated with low-self-esteem individuals.

One final point which can be made here may serve to support further the point of view advanced. An examination of Table 3 indicates that there was generally a greater discrepancy over both "low" and "high" satisfaction conditions for the Low-Self-Esteem group, although the groups had been equated for "satisfaction" and all had been on the job at least 6 mo. Thus, if we ignore the "satisfaction with job" ratings, then we could argue that the probability that one would remain on a job at least 6 mo. is more likely to depend on "personal need satisfaction" for high-self-esteem than for low-self-esteem individuals. While this is a concurrent-validity post hoc type of relationship which needs further investigation, it is strongly consistent with the general proposals advanced.

REFERENCES

BLUMENFIELD, W. Perceived importance and obtainability of certain values associated with naval service. *Psychological Reports*, 1965, 17, 603–606.

CRITES, J. O. Factor analytic definitions of vocational motivation. *Journal of Applied Psychology*, 1961, 45, 330–337.

DAWIS, R. V., ENGLAND, G. W., & LOFQUIST, L. H. *A theory of work adjustment*. Minnesota Studies in Vocational Rehabilitation: XV. University of Minnesota, 1964.

EDWARDS, A. C. *Edwards Personal Preference Schedule*. New York: Psychological Corporation, 1954.

GHISELLI, E. E. The validity of management traits related to occupational level. *Personnel Psychology*, 1963, 16, 109–113.

GHISELLI, E. E. *The Self-Description Inventory Manual*. Available from author, University of California, Berkeley, undated.

KORMAN, A. K. Self-esteem variable in vocational choice. *Journal of Applied Psychology*, 1966, 50, 479–486.

KORMAN, A. K. Self-esteem as a moderator of the relationship between self-perceived abilities and vocational choice. *Journal of Applied Psychology*, 1967, 51, 65–67.

MURRAY, H. A. *Explorations in personality*. New York: Oxford University Press, 1938.

SCHAFFER, R. H. Job satisfaction as related to need satisfaction in work. *Psychological Monographs*, 1953, 67(14, Whole No. 364).

SCHLETZER, V. SVIB as a predictor of job satisfaction. *Journal of Applied Psychology*, 1966, 50, 5–8.

VROOM, V. *Work and motivation*. New York: Wiley, 1964.

(Received January 11, 1967)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 539–543

VOCATIONAL INTERESTS OF COMMUNITY RECREATION ADMINISTRATORS USING THE SVIB

KEITH B. ROYS, Sr.

University of Missouri Department of Recreation and Park Administration

A Community Recreation Administrator scale for the SVIB for Men was developed by testing 433 members of this occupation. The interests of community recreation administrators were most like those of public administrators, YMCA secretaries, and Chamber of Commerce executives and least like physicists, mathematicians, chemists, and engineers.

The SVIB for Men is widely used and differentiates between more than 50 occupations and therefore was selected to determine whether male community recreation administrators could be distinguished from men in other occupations. The resulting scale should aid vocational guidance counselors in helping young people who are considering, or should consider, a career in community recreation.

Many Recreation majors in universities have not been admitted to a Recreation curriculum until their second or third year of college because they did not know that such a program existed. If counselors in high schools and colleges knew of this course of study, advisees interested in community recreation could be apprised of the field early in their academic career.

With the increase in employee recreation programs, this new scoring key could be an asset to personnel directors in business and industry in the selection of new employees or relocation of employees when a move is indicated.

Method

Directories and membership rosters of state recreation societies and associations in the United States were used in compiling a list of 1,000 names and addresses of potentially qualified people. A letter was sent to each, along with a self-addressed return postcard requesting information on sex, current occupation, age, education, and experience. Six hundred twenty-five postcards were returned; 560 were apparently qualified. The 1964 research edition of the SVIB was sent to these 560; 492 or 88% were returned, after only one follow-up letter. Four hundred thirty-three were found to be actually qualified.

A randomly selected criterion group of 350 was compared with the SVIB Men-in-General (M-I-G) group. Each item on which the two groups' responses differed by 17% or more was keyed a plus or minus one depending on the direction of the difference. The remainder of the qualified responses (83) were used as a cross-validation group.

The criterion group was scored with this key. The mean and standard deviation of the raw scores were used to establish standard scores using the traditional formula.

$$\text{Standard Score} = \frac{(M - \bar{X})}{\sigma}\, 10 + 50,$$

where: M = Raw score, $\bar{X}$ = Mean of criterion group, σ = Standard deviation of criterion group.

Letter designations were assigned to standard scores in the same manner usual for SVIB scoring keys; that is, 45 and over = A, 40–44 = B+, 35–39 = B, 30–34 = B−, below 30 = C.

Test-retest reliabilities of the new key are reported in Table 1 for five groups with intertest intervals ranging from 2 wk. to 30 yr. Scores on the key appear to be highly stable ($r = .9$) over periods of 2–4 wk.; moderately stable ($r = .6$–$.7$) over periods of one to two decades.

Validity was established by using the following method. Using Tilton's (1967) table, the percentage

TABLE 1

Reliability of the Community Recreation Administrator Scale Using Test-Retest Correlations

Time between test and retest	r	Group
2 wk.	.93	University of Minnesota sophomores (n = 139)
30 days	.92	Army personnel (n = 102)
8 yr.	.62	Minnesota high school seniors, later University of Minnesota graduates (n = 171)
22 yr.	.72	Stanford University seniors 1929 (n = 191)
30 yr.	.50	Bankers in State of Minnesota (n = 93)

TABLE 2

STANDARD SCORE MEANS, STANDARD DEVIATIONS, AND PERCENTAGE OF OVERLAPS OF THE VALIDATION, CROSS-VALIDATION, AND MEN-IN-GENERAL GROUPS SCORED ON THE COMMUNITY RECREATION ADMINISTRATOR SCALE

	Group		
	Criterion	M-I-G	Cross-Validation
Standard score *M*	50.0	18.4	49.6
SD	10.0	13.4	10.9
Percentage of overlapping	18%		20%

of overlap in the scores of the criterion and M-I-G groups was computed and compared with the percentage of overlap of the cross-validation and M-I-G groups. The criterion group had an 18% overlap with M-I-G; the cross-validation group a 20% overlap—a shrinkage of only 2% (see Table 2).

RESULTS

Scoring the community recreation administrator criterion group members on other available occupational scales, it was found that the occupational group with the greatest similarity of interests to men in recreation was that of public administrator (see Table 3 and Figure 1). However, in scoring the

TABLE 3

HIGHEST AND LOWEST SVIB SCALE MEANS FOR THE COMMUNITY RECREATION ADMINISTRATOR CRITERION GROUP

High		Low	
Scale	Standard score	Scale	Standard score
1. Public Administrator	46.89	1. Physicists	6.85
2. YMCA Secretary	44.38	2. Mathematician	10.65
3. Chamber of Commerce	43.92	3. Chemist	11.90
4. Social Science Teacher	41.85	4. Engineer	13.29
5. Physical Therapist	41.72	5. Carpenter	13.73
6. Social Worker	41.01	6. Architect	15.47
7. Rehabilitation Counselor	40.85	7. Dentist	16.74
8. Credit Manager	39.47	8. Senior CPA	18.69
9. Business Education Teacher	38.27	9. Artist	19.46
10. Personnel Manager	35.68	10. Biologist	20.45

TABLE 4

OCCUPATION GROUPS THAT SCORED HIGHEST AND LOWEST ON COMMUNITY RECREATION ADMINISTRATOR SCALE

High		Low	
Group	Standard score	Group	Standard score
1. YMCA Secretary	45.16	1. Physicists	4.50
2. Chamber of Commerce	39.55	2. Mathematician	6.38
3. Social Worker	34.49	3. Chemist	9.28
4. YMCA Physical Director	34.03	4. Engineer	9.72
5. Business Education Teacher	33.86	5. Artist	10.90
6. Rehabilitation Counselor	33.26	6. Architect	11.46
7. Credit Manager	32.27	7. Dentist	13.94
8. Social Science Teacher	32.21	8. Experimental Psychologist	14.16
9. Physical Therapist	32.07	9. Author-Journalist	14.68
10. Air Force Pilot	30.19	10. Biologist	15.90

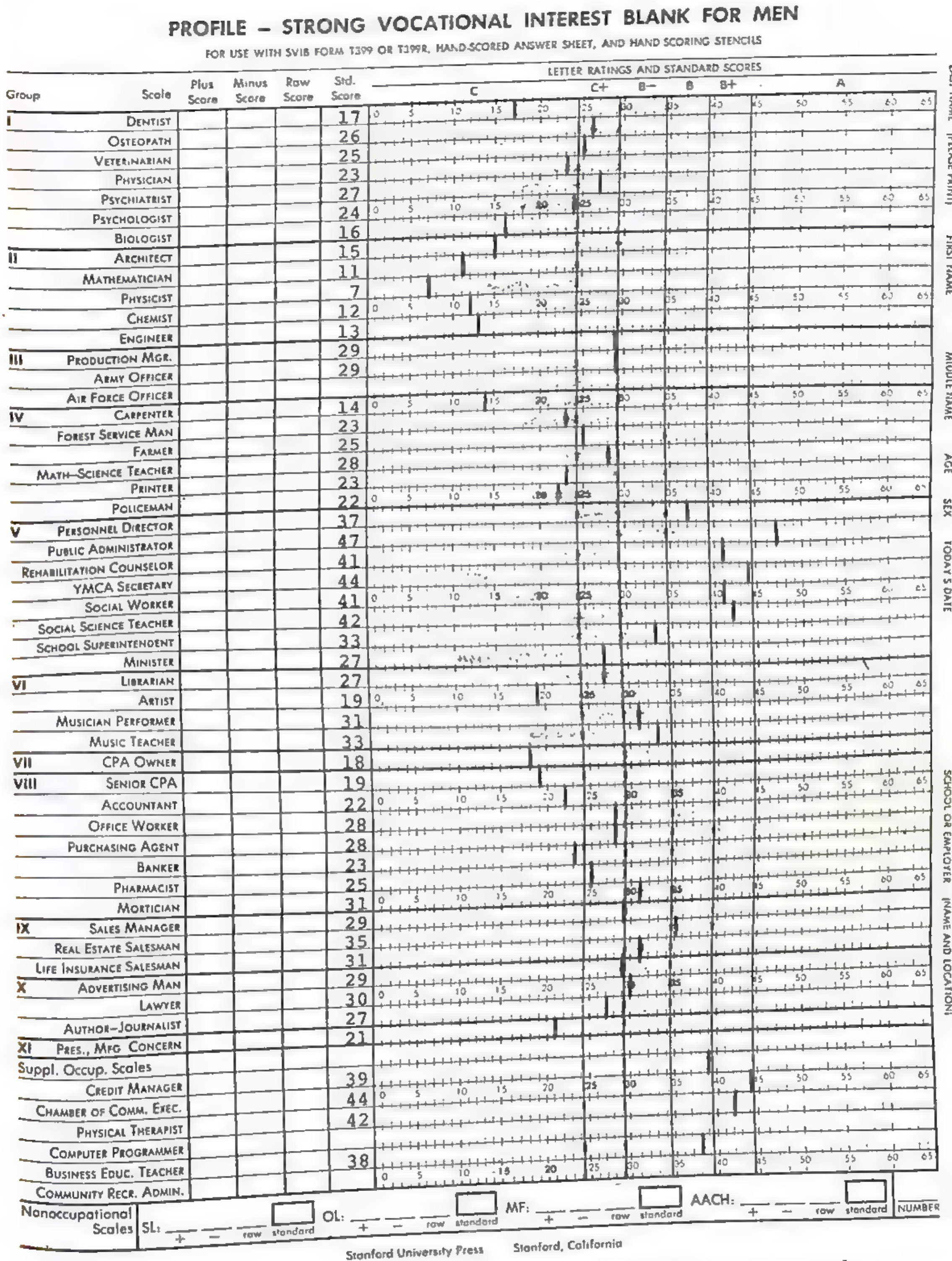


FIG. 1. How the Community Recreation Administrator criterion group scored on other occupation scales.

representatives of the other occupational reference groups on the new community recreation administrator's scale, the converse was not true (see Table 4 and Figure 2). Community recreation administrators had interests in common with many public administrators;

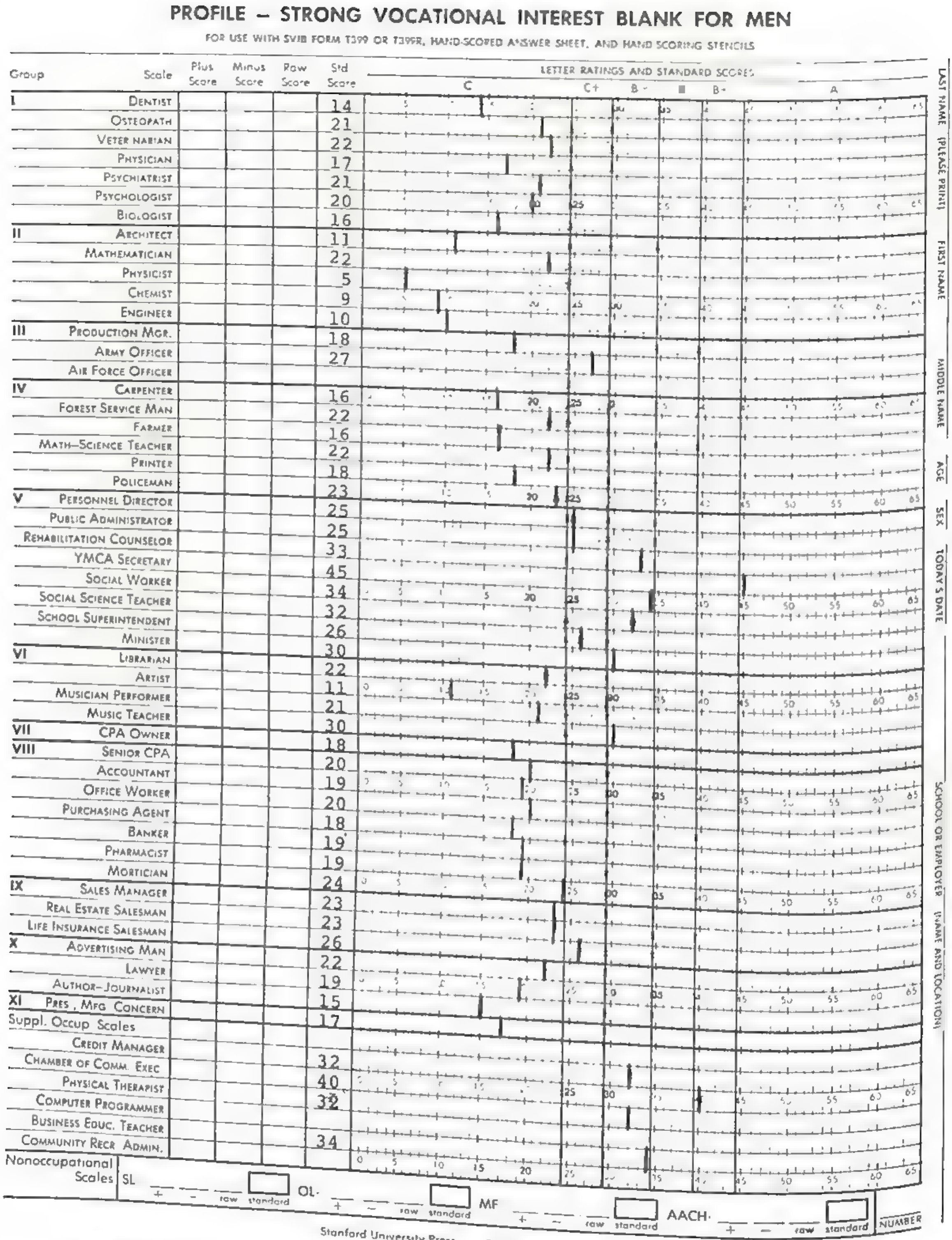


FIG. 2. How other occupations scored on the Community Recreation Administrator scale.

most public administrators, however, did not have the same interests as community recreation administrators. The public administrator key had been derived from a criterion group that included city managers, hospital administrators, public recreation administrators, and various others.

Table 3 also shows that community recrea-

TABLE 5

CORRELATION BETWEEN COMMUNITY RECREATION ADMINISTRATOR SCALE AND OTHER OCCUPATIONAL SCALES ON THE SVIB

Highest		Lowest	
Scale	*r*	Scale	*r*
1. YMCA Secretary	.93	1. Physicist	−.80
2. Chamber of Commerce	.82	2. Architect	−.80
3. Social Worker	.75	3. Engineer	−.78
4. Rehabilitation Counselor	.74	4. Mathematician	−.74
5. Credit Manager	.71	5. Chemist	−.73
6. Business Education Teacher	.71	6. Dentist	−.71
7. Social Science Teacher	.68	7. Artist	−.65
8. Public Administrator	.64	8. Biologist	−.63
9. Personnel Director	.61	9. Physician	−.53
10. Superintendent	.60	10. Farmer	−.53

tion administrators scored high on the YMCA Secretary and the Chamber of Commerce Executive scales. It was interesting to note that the men who were employed in these fields also scored high on the new community recreation administrator's scoring key. Correlation of the scores of the community recreation administrator with the last two vocational groups was also high—.93 and .82, respectively (see Table 5).

TABLE 6

POPULAR INTERESTS INDICATED BY THE COMMUNITY RECREATION ADMINISTRATOR CRITERION GROUP

Item number	%	Item
1. 224 (L)[a]	98.6	Conscientious people
2. 242 (L)	98.0	People who are natural leaders
3. 207 (L)	97.4	Taking responsibility
4. 396 (1)[a]	96.9	Very much interested in my work
5. 236 (L)	96.0	Energetic people
6. 372 (Yes)[a]	95.4	Derive the greatest pleasure and reward from doing a *high-quality* job
7. 344 (1)	94.0	Great variety of work versus similarity in work
8. 127 (L)[a]	93.4	Physical education
9. 222 (L)[a]	91.7	Competitive activities
10. 206 (L)	90.9	Meeting and directing people
11. 243 (L)	90.6	People who assume leadership

Note.—Material in parentheses refers to which column on the SVIB the respondent checked. For example, where an L is marked, the respondent "liked" the item as opposed to "indifference" or "disliked" the item. The "Yes" refers to the fact that 95.4% marked the column which said they "would derive pleasure . . ." as opposed to "would not derive pleasure. . . ." Number 1 refers to the check mark made by the respondent in Column 1 of the SVIB.

[a] 1964 edition of the SVIB only.

TABLE 7

DIFFERENTIAL INTEREST ITEMS ON THE COMMUNITY RECREATION ADMINISTRATOR SCORING KEY

Item number	Differential	Item
1. 70 (L)	54.4	Playground director
2. 7 (L)	53.1	Athletic director
3. 243 (L)	40.6	People who assume leadership
4. 202 (L)	36.9	Teaching children
5. 225 (L)	36.4	Continually changing activities
6. 199 (L)	33.7	Making speeches
7. 131 (L)	33.1	Public speaking
8. 331 (3)	33.1	Deal with people
9. 158 (L)	32.9	Conventions
10. 196 (L)	31.7	Interviewing men for a job

Note.—See Note of Table 6.

A lack of mutual interests between community recreation administrators and physicists, mathematicians, chemists, and engineers was demonstrated by the low standard scores seen in Tables 3 and 4 and by the negative correlations seen in Table 5.

The popular responses of the community recreation administrator criterion group members are listed in Table 6. These items fell into three categories: the first was related to the personality of the individual—they liked conscientious people who were natural leaders displaying energy and assuming leadership. A second category reflected something of their personal activity interests—a desire for competitive activities and physical education. The third category related to their work—it indicated that they were interested in and enjoyed their work, and received a great deal of pleasure and reward from doing a large variety of things well.

Differential interests (see Table 7) showed the man in recreation enjoyed working actively with people, for example, playground directors, athletic directors, and teachers of children. He was also concerned with factors other than face-to-face leadership in active settings, as indicated by the high percentages on the items which related to administrative responsibilities such as interviewing men for jobs.

REFERENCES

TILTON, J. W. The measurement of overlapping. *Journal of Educational Psychology*, 1937, 28, 656–662.

(Received January 16, 1967)

Journal of Applied Psychology
1967, Vol. 51, No. 6, 544–546

PREDICTIVE VALIDITY OF A BIOGRAPHICAL INVENTORY IN HIGHER EDUCATION

RICHARD S. MACKENZIE

University of Pittsburgh, School of Dentistry

Scores on the Extracurricular Achievement Record, developed by Holland and Nichols, predicted professional extracurricular activities of 102 dental students. These activities involved leadership, simulated professional meetings, and research. 3 hypotheses were tested: (1) students seeking candidacy for class office in the 1st-yr. class election have higher scores than noncandidates on the Leadership scale of the Extracurricular Achievement Record ($p < .05$); (2) students giving table clinics at the Student American Dental Association Day have higher scores than nonparticipants on the Total scale of the Extracurricular Achievement Record ($p < .001$); (3) students applying for summer research fellowships have higher scores than nonapplicants on the Scientific scale of the Extracurricular Achievement Record ($p < .001$). These hypotheses each supported the predictive validity of the biographical inventory.

In 1964, Holland and Nichols developed a college-level biographical inventory of extracurricular accomplishment called the Extracurricular Achievement Record. Included in this inventory were subscales for measuring achievement in science, dramatic arts, literature, music, art, and leadership. The following are examples taken from the Science subscale.

Check those achievements that apply to you.

_____ Built a piece of equipment or laboratory apparatus on my own. (not as part of a course) Describe briefly.

_____ Appointed a teaching assistant in a scientific field.

_____ Received a research grant. Indicate nature.

The entire list of statements included in both the Scientific subscale and the other subscales can be found in Holland and Nichols (1964).

Among the personality measures, biographical inventories are considered to be the most reliable predictors of future performance (Nunnally, 1959). The biographical inventory has been used successfully to predict pilot performance (Guilford, 1947) and salesman success (Ferguson, 1952; Sorenson, 1965). Because of this evidence, it was reasoned that the Extracurricular Achievement Record, as an inventory of preprofessional collegiate extracurricular activities, might be valuable as an aid in selecting those students for dental school who would take an active interest in professional activities. The intent of this study was to determine whether the Extracurricular Achievement Record does, in fact, have predictive validity for professional activities in dental school.

Method

Subjects. The *Ss* were members of the first-year class at the University of Pittsburgh School of Dentistry. In the fall of 1965, this class consisted of 104 students from 51 undergraduate colleges.

Predictive measure. The Extracurricular Achievement Record of Holland and Nichols was the predictive measure used in this study. It was administered to 102 students during Orientation Week in September 1965. The score on the Total scale and each subscale was determined by adding the number of items checked.

Criterion measures. The three criterion measures of professional activities used to test the predictive validity of the Extracurricular Achievement Record were: (1) student candidacy in class election; (2) student participation in giving table clinics in the Student American Dental Association (ADA) Day; (3) student application for summer research fellowships.

Procedures. Each of the three criterion measures divided the students into two groups, participants and nonparticipants. The scores made earlier by the two groups on the Extracurricular Achievement Record were compared to test the following three hypotheses: (1) students seeking candidacy for class office in the first-year class election have higher scores than noncandidates on the Leadership scale of the Extracurricular Achievement Record; (2) students giving table clinics at the Student ADA Day have higher scores than nonparticipants on the Total scale of the Extracurricular Achievement Record; (3) students applying for summer research fellowships have higher scores than nonapplicants on the Scientific scale of the Extracurricular Achievement Record.

Hypothesis I. Two weeks after the semester began, the first-year class decided that candidates for class office must each give a 2-min. speech indicating his reasons for desiring candidacy. The aspiring student leaders were given 1 wk. to make their decisions, prepare their speeches, and gather electoral support. Of the 102 possible candidates, the names of 18 students appeared on the mimeographed ballot as candidates for office. It was reasoned that students who had high scores on the Leadership scale would be more likely to seek elective office than those who did not have high scores.

Hypothesis II. On March 16, 1966, the local Student ADA held their fifteenth annual meeting. This meeting is modeled after State and National Dental Meetings and provides an opportunity for the students to develop the habit of participating in organized professional activities. The table clinics put on exclusively by students provide experience in demonstrating new techniques, discussing dental problems, or presenting a research project. The table clinics are given in separate booths and consist of a formal presentation followed by questions and answers. Participation is entirely voluntary, just as it is in the professional organization. Table clinics are presented by 19 first-year students out of 96 (6 were lost at midyear). It was reasoned that students who had high scores on the Total scale for extracurricular achievement would be more likely to participate in professional-type activities than those who did not have high scores.

Hypothesis III. Each summer various institutions such as Veterans Administration Hospitals, the Public Health Service, and various colleges and universities offer opportunities for dental students to participate in research activities. There is competition for these positions and interested students must apply early. On April 12, 1966, the first-year students were given a questionnaire asking what type of summer employment they were planning and whether they had applied for a research fellowship. If they had applied for a research fellowship, they were asked to describe the position. Research fellowships were applied for by 17 out of 96 students. It was reasoned that those students who had high scores on the Scientific scale would be more likely to apply for summer research fellowships than those who did not have high scores.

RESULTS

Hypothesis I. Students seeking candidacy for class office in the first-year class election did score higher than noncandidates on the Leadership scale of the Extracurricular Achievement Record. The scores that the 18 candidates for class office made on the Leadership scale were compared to the scores of the remaining 86 students. The 18 candidates had checked 53 out of a possible 216 leadership items. The 84 noncandidates had checked 181 items out of a possible 1,008 leadership items. A

TABLE 1

COMPARISON OF SCORES ON THE LEADERSHIP SCALE FOR CANDIDATES AND NONCANDIDATES FOR CLASS OFFICE

For class office	Leadership items		Total
	Checked	Not checked	
Candidate	53	163	216
Noncandidate	181	827	1008
Total	234	990	1224

Note.—$p < .05$. $\chi^2 = 5.26$ with 1 *df*.

chi-square analysis was done on the above 2×2 table.

The null hypothesis was that the distribution of scores for the candidates and noncandidates should not differ from the marginal totals more than would be expected on the basis of chance. This null hypothesis was rejected at the $p < .05$ level. Thus, scoring high on the Leadership scale predicted participation in leadership activities.

Hypothesis II. Students giving table clinics on Student ADA Day did score higher than nonparticipants on all scales of the Extracurricular Achievement Record. The college extracurricular achievement scores of the 19 students giving table clinics were compared to the scores of the 77 remaining students. The participating students had checked 145 out of a possible 1,083 items. The nonparticipating students had checked 404 out of a possible 4,389 items. A chi-square analysis was done on the following 2×2 table.

The null hypothesis was that the distribution of scores for the participants and non-

TABLE 2

COMPARISON OF SCORES ON THE EXTRACURRICULAR ACHIEVEMENT RECORD FOR PARTICIPANTS AND NONPARTICIPANTS IN A STUDENT ADA DAY

Student ADA Day	Total items		Total
	Checked	Not checked	
Participant	145	938	1083
Nonparticipant	404	3985	4389
Total	549	4923	5472

Note.—$p < .001$. $\chi^2 = 17.2$ with 1 *df*.

TABLE 3

COMPARISON OF SCORES ON THE SCIENTIFIC SCALE FOR APPLICANTS AND NONAPPLICANTS FOR SUMMER RESEARCH FELLOWSHIPS

For research fellowship	Scientific items		Total
	Checked	Not checked	
Applicant	59	179	238
Nonapplicant	142	950	1092
Total	201	1129	1330

Note.—$p < .001$. $\chi^2 = 21.1$ with 1 *df*.

participants should not differ from the marginal totals more than would be expected on the basis of chance. This null hypothesis was rejected at the $p < .001$ level. Thus, scoring high on the Extracurricular Achievement Record also predicted participation in extracurricular professional activities.

Hypothesis III. Students applying for summer research fellowships did score higher than nonapplicants on the Scientific scale of the Extracurricular Achievement Record. The scores that the 17 applicants for summer research fellowships made on the Scientific scale of the inventory were compared to the scores of the remaining 78 students. The 17 applicants had checked 59 out of a possible 238 items of the Scientific scale. The 78 nonapplicants had checked 142 out of 1,092 possible scientific items. A chi-square analysis was done on the above 2 × 2 table.

The null hypothesis was that the distribution of scores for the applicants and nonapplicants should not differ from the marginal totals more than would be expected on the basis of chance. This null hypothesis was rejected at the $p < .001$ level. Thus, scoring high on the Scientific scale predicted a tendency to seek research positions.

DISCUSSION

The findings of this study have supported the predictive validity of the Extracurricular Achievement Record of Holland and Nichols. Students who participated in leadership, scientific, or general extracurricular activities as undergraduates continued to participate in similar activities while in dental school.

Because behavioral patterns tend to remain stable over time, it is reasonable to expect not only that an active undergraduate will participate in activities as a dental student, but also that he will participate in various professional and community functions thereafter as a practicing dentist. Since most professions are interested in members who have leadership potential and scientific and cultural interests, it is assumed that educators in the professions would be interested in utilizing biographical inventories such as the Extracurricular Achievement Record in selecting candidates for their schools.

From a training point of view, if students could be guided by various instructional strategies in professional schools to develop leadership skills, scientific inclinations, and the habit of participating in professional activities, then a measure similar to the Extracurricular Achievement Record might be utilized to measure the results. For example, if a group of students was effectively trained in leadership skills and another group of students was not, then the experimental group should score higher on a biographical inventory concerning leadership activities following graduation.

REFERENCES

FERGUSON, L. W. *Personality measurement.* New York: McGraw-Hill, 1952.

GUILFORD, J. P. (Ed.) *Printed classification tests.* Washington, D. C.: Government Printing Office, 1947.

HOLLAND, J. L., & NICHOLS, R. C. Prediction of academic and extracurricular achievement in college. *Journal of Educational Psychology*, 1964, **55**, 55–65.

NUNNALLY, J. C. *Tests and measurements.* New York: McGraw-Hill, 1959.

SORENSON, W. W. Configural scoring of biographical items for predicting sales success. *Dissertation Abstracts*, 1965, **25**, 5379–5380.

(Received February 16, 1967)

journal of applied psychology monograph

Vol. 51, No. 4
Part 2 of 2 Parts

August 1967
Whole No. 636

A PSYCHOLOGICAL STUDY OF PAY[1]

MASON HAIRE, E. E. GHISELLI AND M. E. GORDON

University of California, Berkeley

Life Insurance Agency Management Association, Hartford, Connecticut

A study in 3 sections of empirical data on managerial compensation for 3 groups of managers (N = about 90 for each group): (1) a description of the distributional characteristics of pay over time (about 25 yr.); (2) the correlation of pay with pay over time and presents statistical analyses to explain the observed relationships; (3) a statistical model of pay capitalizing on the cumulative character of pay (pay at Year n is composed of pay at Year 1 + raises at Years 2, 3, . . . n) and the formulas for the distributional and correlational character of composites. The potential psychological leverage of hitherto little-considered variables stands out—for instance, without increasing the total salary bill, management of the variance of pay over a group in a given year and of the correlation of pay with raises from year to year allows one to deal with the level of aspiration of the individual and his relative standing in the group. The managerial implications of the statistical behavior of pay are discussed in detail.

1. Distribution of Pay over Time

It is a truism to say that one knows very little about the psychological variables associated with pay. Even so, it needs to be said and repeated. On the one hand, on the side of psychology, pay partakes of a variety of crucial theoretical issues. For example, although it is often sneered at as a motivator, even psychologists, deep down, probably have an implicit faith in it. A variety of studies deal with it as a source of satisfaction (Herzberg, Mausner, & Snyderman, 1959; Lawler & Porter, 1963). Theoretical frameworks include it under a variety of guises (Vroom, 1964). As an implement of variable reinforcement schedules, pay makes contact with a whole field of psychology. Yet, except for a group of pious generalizations, precious little is written about it. On the other hand, from the side of industry, there is also a variety of compelling reasons for attention. Pay is widely accepted as a prime motivator. Very large sums are discretionarily committed on the basis of assumptions about its action. Large amounts of so-called incentive pay are spent without any evidence about who becomes incensed, at what point, for how long, or under what circumstances.

There is relatively little data on the behavior of pay. Companies and professional organizations frequently collect cross-sectional data—so-called maturity curves—showing, for example, average earnings since the highest degree (Patton, 1961; Weissinger, 1958). For many psychological purposes, these have the usual drawbacks of cross-sectional data. It is impossible to say what happened to anyone over time in the data, and the measured money spent becomes the dependent variable rather than its impact on an individual. These data are useful to companies, since they provide

[1] This research was carried out with the help of a grant from the General Electric Company and the support and facilities of the Institute of Industrial Relations, University of California, Berkeley. Important contributions were made by a variety of people, especially M. E. Gordon, University of California, Berkeley.

guidelines for following either an implicit or explicit policy of paying a rate characteristic of the industry and the neighborhood. However, this very kind of imitation in pay policy has inhibited a creative use of the motivational leverage presumably inherent in pay. To support and guide this kind of creative rethinking of the psychological characteristics of pay, various kinds of data are needed. This monograph will present three facets of compensation which seem particularly relevant to a psychological analysis: characteristics of the distribution of pay over a period of time, the correlations of pay with pay over time, and, finally, a statistical model of compensation. Included are three longitudinal studies of the pay histories of particular individuals. As far as is known, only one report on pay has been conducted on the basis of a longitudinal research design—a recent publication by Brenner and Lockwood (1965).

As one turns to longitudinal data, several different parameters stand out, each with some possible psychological meaning. It may be something of an oversimplification to deal with them in the following manner, but it will get us into the problem. From the side of the company, there is, first, the change in average pay over time. This is, of course, importantly, a variable cost to the company. In motivational terms, at any one time it represents the company's bid in the labor market and its evaluation of the services required. Over time, the change in the average pay represents a recognition of the rate of development of the group, of their usefulness to the company, and, if one assumes perfect mobility, of their usefulness to someone else.

The standard deviation of pay, on the other hand, would seem to reflect organizational variables; that is, the amount of difference in the job assignments of the group largely determines the spread of pay. This is complicated by the fact that individual differences will be recognized by variable raises. Since pay is importantly a cumulative affair, that is, pay at Year n is equal to pay at Year 1 + raises on Years 1, 2, . . . $n-1$, the standard deviation of increments influences the standard deviation of pay. The standard deviation of pay is similarly a composite. The variation in pay on a given year is a function of the variation in pay the year before plus the variation in the raises between the 2 years. This is further complicated by the fact that, because of the covariance term in the formula for the standard deviation of composites ($\sigma_{k+i} = \sigma_k + \sigma_i + 2\sigma_k\sigma_i r_{ki}$), the influence of σ_i (the standard deviation of increments) on σ_k (the standard deviation of pay) depends partly on the correlation between increments and pay. Thus there is another parameter in longitudinal data, r_{ki}, or "who gets the raises?"

At this point four parameters, from the side of the company, have been identified, which have, at least in simple terms, considerable policy implications:

1. $\bar{k}$, the average pay, which represents at any one time the company's bid for services, and over time, its assumptions about the role of development.

2. σ_k, the standard deviation of pay, which measures the differences in the value of individual assignments in the group, and hence is primarily an organizational variable.

3. σ_i, the standard deviation of raises, which represents the company's assumption about the size of the distribution of individual differences in the group at any one time.

4. r_{ki}, the correlation between pay and raises which may reflect the operation of the company's performance appraisal program over time, and, in any case, involves assumptions about the consistency of goodness or usefulness of performance.

These variables all have important weight in a company's policy about compensation, and they are complexly interrelated. Subsequent data will show something about how they behave in fact, and some theoretical models of their relationships will illustrate their role in possible pay policies. First, look briefly at these variables from the side of the individual.

First, $\bar{k}$, to the individual, means the value of his services in the market. Over time it means the rate of his growth and development. More than that, since it is money we are talking about, it is related to what he gets and what he needs. Since most companies pay on some form of a percentage increase over time, the plot of $\bar{k}$ is typically positively accelerated. The individual's consumption function has a quite different shape, and this discrepancy has im-

portant motivational implications. In passing, it might be pointed out that most people, when questioned, do not seem to recognize that they are actually on this positively accelerated curve of salary. This datum, which will not be gone into here, points to the important area of how pay is perceived—a psychological problem which must be explored much further before the impact of pay on the individual will be understood.

σ_k, to the individual, gives him a measure of the range of possibilities, and, at any one time, limits his (local) aspirations. σ_i is a more direct index of the evaluation of specific efforts or accomplishments. It is probably here, if anywhere, that an incentive will appear. Finally, r_{ki}, from the point of view of the individual, treats the crucial question "Who gets the raise?" It can vary from "The one who got paid most before" ($r_{ki} = 1.0$) through "It will be distributed at random" ($r_{ki} = 0$) to "The one who got paid least before" ($r_{ki} = -1.0$). The motivational implications of these possibilities are clear. One should also notice that the stability of r_{ki} over time may be an important psychological variable.

Empirical Data

The first job that needed to be done was to collect some simple descriptive data about what happens to pay on the basis of which psychological interpretations can be made. Specifically, the means, standard deviations, and other parameters of the distributions of both salaries and raises were examined for each year in a 25-yr. period. The effects of various changes in these parameters will be discussed from the viewpoint of the individual and the organization.

Three samples of executives were used in this study. These samples have been identified as Companies A, B, and X. Companies A and B represent two groups of executives employed by the same large, diversified, manufacturing corporation. From Company A, 90 executives were selected on the basis of whether they had reached Level 20 or above on the company pay scale[2] by January 1, 1963, and had at least 20 yr. of service with the company. Company B subjects (*S*s) were selected from Levels 15–17 on the company scale.[3] These individuals were approximately matched with the *S*s in Company A on the basis of age and years of service. Company X is a large manufacturing corporation in a process industry. From Company X, 90 executives were selected in a quasi-random manner by choosing every *n*th man from among all employees whose service record extended beyond 15 yr. with the company and who earned more than $8,000 a year at the time of the study.

The salary paid to each of the *S*s for each year of employment was provided by the companies. The raw data were translated to annual salaries from the dates of raises in the earning history of each executive. Bonuses, stock options, and other forms of compensation were not included in the salary figures. Anonymity was preserved for all *S*s involved in the project, the research team having access to only the salary data and the service dates of each of the individuals being studied.

A number of studies which have discussed salary in relation to time have attempted to correct the raw pay data for the effects of inflation and increased cost of living. Jaques (1961), for example, used the wages index, which is an index of the average national minimum wage for a normal working week, to eliminate the effects of inflation. By correcting his earnings data to a chosen period with this index, he felt that wages for different calendar years were made comparable.

After preliminary efforts to adjust the salary of a subset of the samples used here for changes in the cost of living, this procedure was abandoned because the adjustments did not substantially change the results obtained. First, the relationship of the parameters over time was not measurably altered, the shape of the plot of the mean salary across calendar years being unaffected by these corrections. Second, the correlations between salary on different years remained about the same for the adjusted salaries as for the unadjusted salaries. Finally, the selection of any particular index for the purpose of adjusting the salaries would only have added a dimension of arbitrariness to the results. Since there is no standard accounting procedure for making this type of correction, but rather a number of such procedures, all of which produce unique results, the salary data reported have not been modified to compensate for changes in the cost of living.

Notation

The following is a sample of the notations and the statistics to be used in the analysis of the data for the *s* *S*s in each of the *c* companies. Note that a dot used as a subscript denotes the fact that summation has occurred over the factor usually represented by a letter occupying the particular position indicated by the dot. Also, a line above a symbol indicates that the particular notation represents an arithmetic mean.

k_{sc}—pay for *S* *s* in Company *c* on Year *t*.
i_{sc}—increment (or decrement) in pay for *S* *s* in Company *c* on Year *t*.
$(k + i)_{sc}$—pay for *S* *s* in Company *c* on Year $t + 1$.
$\bar{k}_{.c}$—average pay for *S*s in Company *c* on Year *t*.
$\bar{k}_{..}$—average pay for all *S*s in all companies on Year *t*.
$\bar{\imath}_{.c}$—average increment in pay for *S*s in Company *c* on Year *t*.
$\bar{\imath}_{..}$—average increment in pay for all *S*s in all companies on Year *t*.

[2] Level 20 of the Company A pay scale corresponds to an annual income of about $40,000.

[3] Levels 15–17 on the Company B pay scale correspond to annual incomes ranging between $20,000 and $22,000.

$(k+i)_{.c}$—average pay for Ss in Company c on Year $t+1$.
$\sigma k_{.c}$—standard deviation (SD) of pay for Ss in Company c on Year t.
$\sigma_{i.c}$—SD of increments in pay for Ss in Company c on Year t.
$\sigma_{(k+i).c}$—SD of pay for Ss in Company c on Year $t+1$.
$\sigma_{(k+i)..}$—SD of pay for all Ss in all companies on Year $t+1$.

TABLE 1

NUMBER OF CASES IN THE UNSELECTED SAMPLE

Year	Company A	Company B	Company X
1919	1		
1920	1		2
1921	1		2
1922	3	1	2
1923	5	2	2
1924	7	2	3
1925	7	6	4
1926	9	8	5
1927	11	11	8
1928	17	16	9
1929	19	19	10
1930	24	23	16
1931	33	31	20
1932	40	33	22
1933	43	36	27
1934	46	39	28
1935	51	45	34
1936	60	54	39
1937	70	66	43
1938	74	71	49
1939	77	75	54
1940	82	76	62
1941	83	83	72
1942	87	88	75
1943	89	89	74
1944	89	89	72
1945	90	89	75
1946	90	89	88
1947	90	89	89
1948	90	89	89
1949	90	89	90
1950	90	89	90
1951	90	89	89
1952	90	89	89
1953	90	89	90
1954	90	89	90
1955	89	89	90
1956	88	89	90
1957	88	88	90
1958	86	86	90
1959	86	84	90
1960	80	82	90
1961	72	69	87
1962	56	52	82
1963	21	29	74

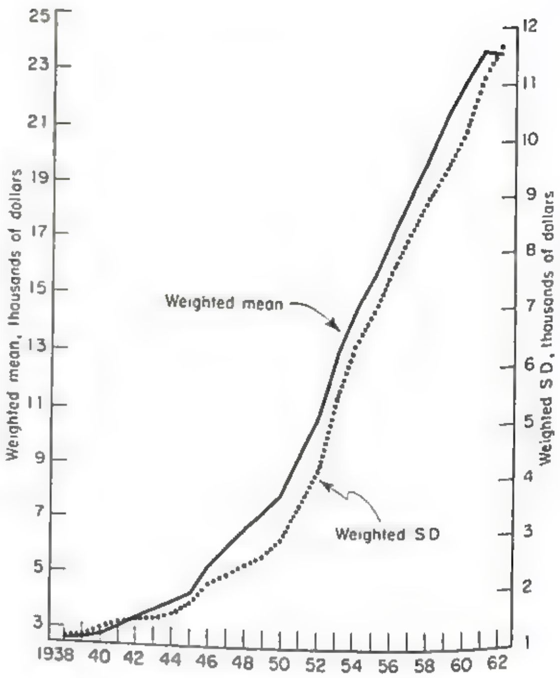


FIG. 1. Weighted mean and standard deviation of salary.

Because of the longitudinal nature of the data, the sample size (N) varied from year to year. The fluctuation in N was caused by different starting and terminating service dates with a company for the different members of a given sample. Although salaries were available for as many as 45 yr., dating back to 1919, the reliability of statistics based upon much of the data, especially from the earlier calendar years, would be variable due to the small N characteristic of this period. A criterion of a minimum of 40 cases in each company on a given calendar year was adopted arbitrarily. This multiple cutoff for the sample size was selected in order that summary measures calculated on the data across companies for a given year would not be unduly influenced by the variance of the scores from a single company. This restriction reduced the data available for analysis to those for the calendar years 1938 through 1962.

Table 1 provides a list of the sample size for all calendar years on which data were available for each company.

Results

Figure 1 contains a plot of $\bar{k}_{..}$ and $\sigma_{k..}$, the weighted mean and standard deviation for all the Ss in the three companies, for the calendar years 1938 to 1962.[4] Figure 2 contains individual graphs of $\bar{k}_{.c}$ for each of the separate

[4] The formula utilized for calculating the standard deviation of the combined samples from the individual company standard deviations can be found in Peatman (1963, p. 64).

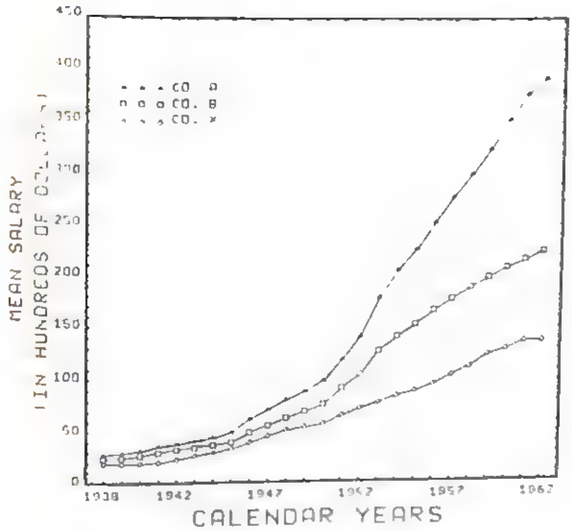


Fig. 2. Average salaries by companies.

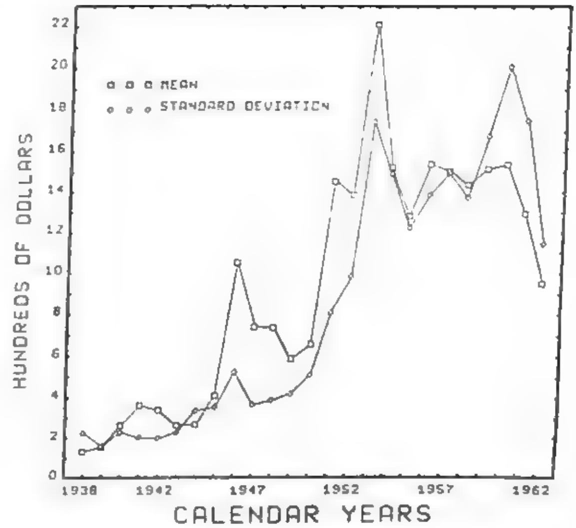


Fig. 4. Weighted mean and standard deviation of increments in pay for combined sample.

companies, while Figure 3 contains individual graphs of $\sigma_{k_{.c}}$ for each company. Notice that the graphs of the average salary are all positively accelerated increasing monotonic functions. The graphs of $\sigma_{k_{.c}}$, though less similar in appearance than those for $\bar{k}_{.c}$, show generally increasing functions except for some minor fluctuations which disappear in the plot of the combined samples ($\sigma_{k_{..}}$). The fact that the curves of $\sigma_{k_{.c}}$ all turn down at the end is partly a function of shrinkage in the sample.

The graphs of $\bar{k}_{.c}$ indicate that on the average the contribution or value of the executive to the organization, as reflected by the magnitude of his salary, increased logarithmically. From the standpoint of the company, the shape of the curves means a policy of something approaching a constant rate of increase in salary, allowing the absolute amount of increases to change logarithmically.

This logarithmic trend in the magnitude of the increments is displayed in Figure 4, which is a plot of the weighted mean and standard deviation of increments based upon the combined samples. The sharp drop in magnitude of these parameters for the last two calendar years is probably an artifact due to the retirement of a number of executives. (The effect of losing data from the sample at the end of an

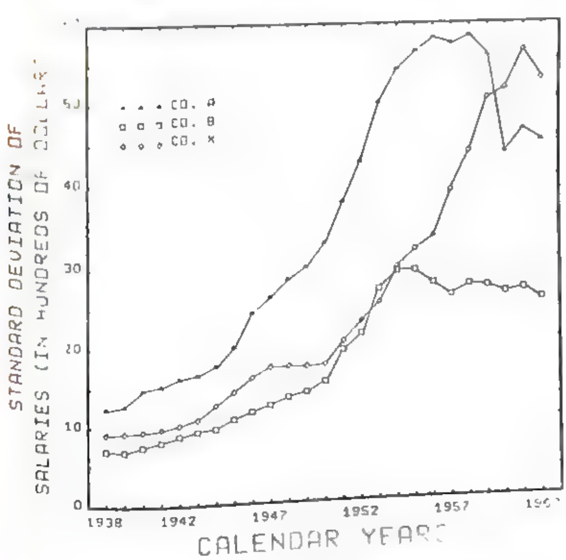


Fig. 3. Standard deviation of pay by companies.

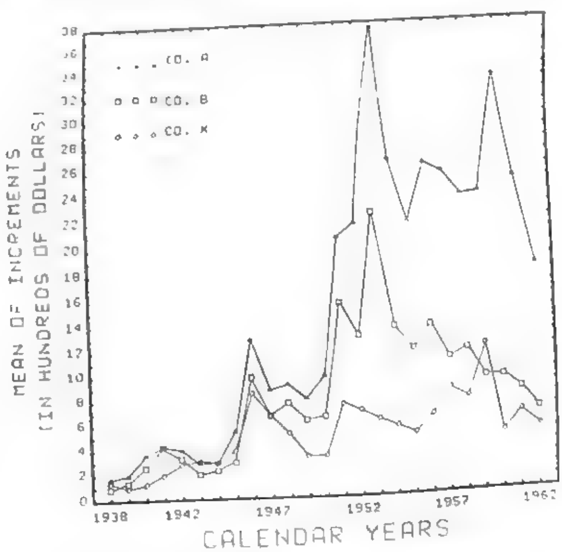


Fig. 5. Average increment in pay by company

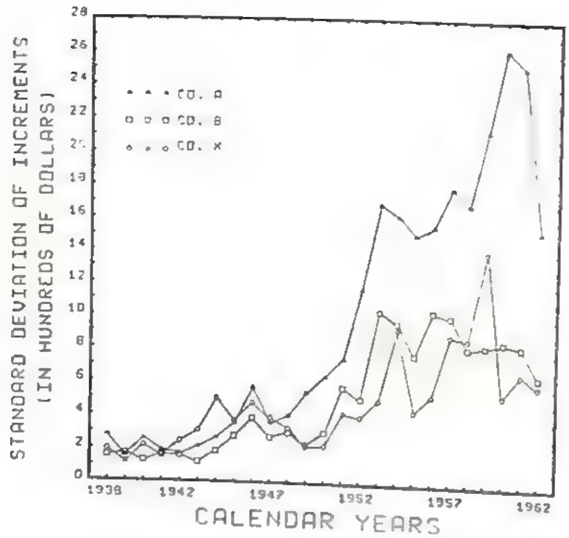


FIG. 6. Standard deviation of increments in pay by companies.

S's career is a marked effect on the average increment and the dispersion of increments when the progression of raises is geometric.) The same phenomenon appeared in Figure 3.

The behavior of raises is further portrayed in Figures 5 and 6, which describe $\bar{\imath}_{.c}$ and $\sigma_{i_{.c}}$ as a function of calendar years. The pattern of raises over time is quite similar for the three companies, although the magnitudes differ considerably among the groups.

The relationship between $\bar{k}_{.c}$ and $\bar{\imath}_{.c}$ is graphically demonstrated in Figure 7. When one considers pay on a given year as pay on the previous year plus an increment, and when the sample size on successive years remains constant, $\bar{k}_{.c}$ on Year n equals the sum of $\bar{k}_{.c}$ and $\bar{\imath}_{.c}$ for Year $n - 1$. That is, one can consider pay on any year as representing starting salary plus the cumulative increments awarded over time. The change in $\bar{k}_{.c}$ from year to year when N is constant is caused entirely by changes in $\bar{\imath}_{.c}$. If there is a fluctuation in N, arithmetic means will not be comparable. Figure 7 is a plot of $\bar{k}_{.A}$ for the calendar years 1945 through 1954. This period was arbitrarily chosen because the sample size remained constant over this 10-yr. span ($N = 90$). The curve of the cumulative $\bar{\imath}_{.A}$—that is, the $\bar{\imath}_{.A}$ increments summed each year across all preceding years—is also plotted. It is immediately apparent that the shape of the two curves is identical, although they are represented on different portions of the dollars dimension. Salary at Year n is the sum of all raises plus a constant (the starting salary).

Figure 8 contains a plot of the index of skewness of the distribution of salaries in each of the companies for each of the 25 yr. studied. Figure 9 contains a plot of the skewness of the distribution of increments. Figure 8 reveals that the distribution of salaries was positively skewed every year in Company X, almost every year in Company A, and three-quarters of the time in Company B. This means that a few individuals received salaries far above the

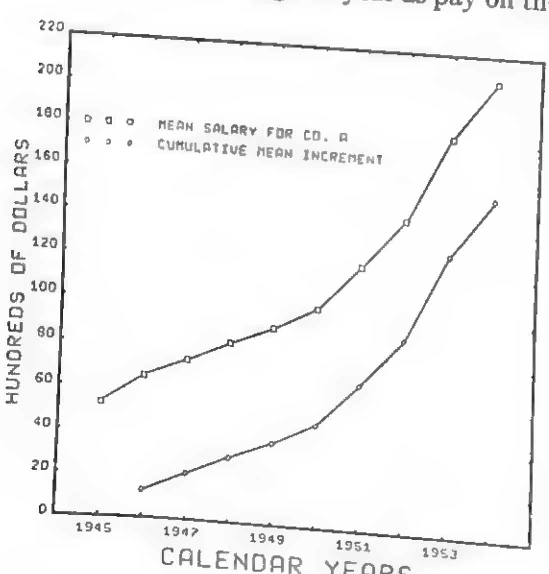


FIG. 7. Average pay and average increment, Company A, 1945–1954.

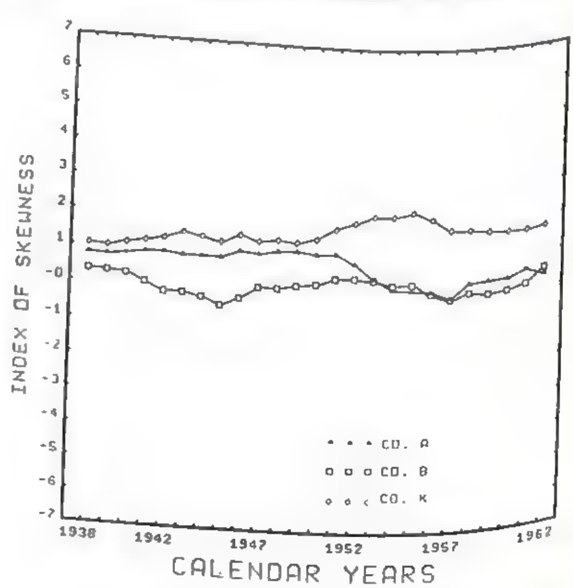


FIG. 8. Skewness of the distributions of salaries by companies.

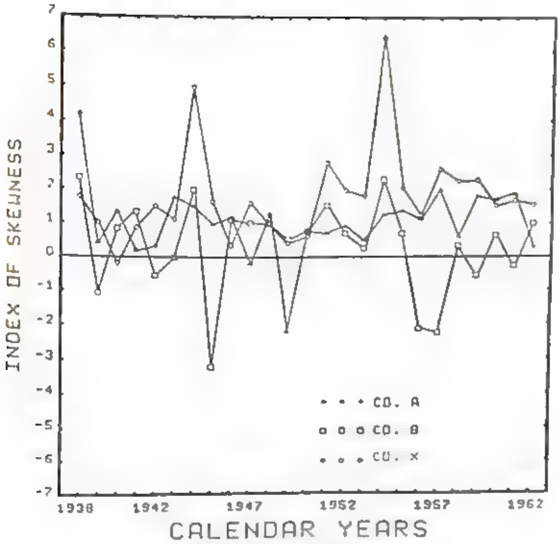


FIG. 9. Skewness of the distribution of increments by companies.

salaries paid to the majority of the executives in the group, especially in Company X, in which departures from normality are uniformly large.

The high degree of skewness was expected in Company A and Company X because there was no ceiling on the salaries represented in these groups. The positively skewed distributions in these two companies are reflections of the pyramidal structure at the top of the organization; that is, skewness was based upon salaries across levels of the organization. In Company B, however, limits on both the minimum and maximum salary were imposed, thereby restricting the range of salaries. In this case, skewness is a reflection of the distribution of intralevel pay.

Because of the high degree of skewness present in the raw salary data, the arithmetic mean, $\bar{k}_{.c}$, cannot be considered alone as an accurate reflection of central tendency. The median was calculated for the distributions of salaries for each year. Figure 10 contains a plot of the medians for each company. The shape of these curves of median salary is quite similar in form to those displayed in Figure 2. Some difference appears in the plot for Company X. Although the curve for the medians of Company X still shows a decided upward trend, the linear component seems to have increased after the effect of the high degree of skewness of the distributions has been removed.

The reason for this flattening of the curve can be explained by the fact that the later years in Company X, that is, those subsequent to 1953, were characterized by the most highly skewed distributions of salary for all companies. The effect of these positively skewed distributions was to elevate the value of the arithmetic mean for these years, which thereby imparted the curvilinear upswing in the graph for Company X over the most recent years. This effect was removed by utilizing the median which produced the largest adjustments in central tendency for those distributions which were most highly skewed.

The plots of both the medians and arithmetic means show that the highest average salary was paid by Company A, the second highest by Company B, and the lowest by Company X; that is, the companies are ordered the same way on the basis of average salary regardless of which measure of central tendency is used. Since *S*s were selected on the criterion of relative salary for inclusion into Companies A and B, it was expected that the average salary for Company A would be higher than that of Company B for the year in which the salary criterion was measured (1963). However, Figures 2 and 10 both indicate that the average salary for Company A was higher than that of Company B throughout the 25-yr. period studied. Furthermore, the difference in the average salaries of these companies progressively increased over time beginning with 1938.

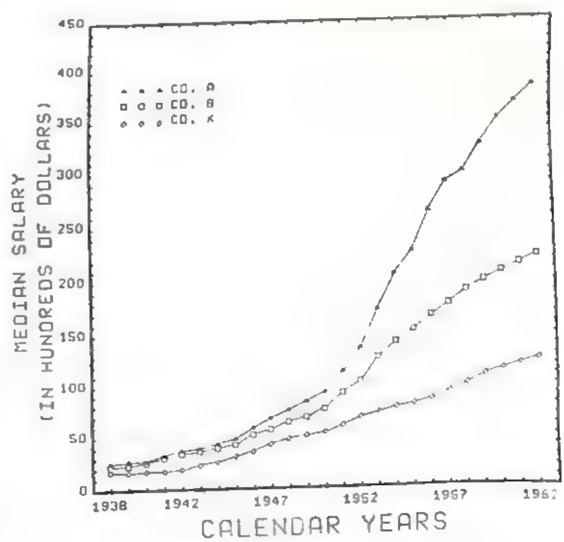


FIG. 10. Median salaries by company.

TABLE 2

CORRELATIONS BETWEEN MEANS AND STANDARD DEVIATIONS FOR TOTAL SAMPLE AND EACH INDIVIDUAL COMPANY

*M*s and *SD*s	*r*
$\bar{k}_{..}$ and $\bar{\sigma}_{k..}$	.996
$\bar{k}_{.A}$ and $\sigma_{k.A}$	.844
$\bar{k}_{.B}$ and $\sigma_{k.B}$	.929
$\bar{k}_{.X}$ and $\sigma_{k.X}$	.984
$\bar{i}_{..}$ and $\bar{\sigma}_{i..}$	.900
$\bar{i}_{.A}$ and $\sigma_{i.A}$	.905
$\bar{i}_{.B}$ and $\sigma_{i.B}$	.814
$\bar{i}_{.X}$ and $\sigma_{i.X}$	.817

In terms of career pay, one might consider the curves of the means of yearly salary as constituting separate paths for the two groups toward the criterion of pay in 1963. However, when the standard deviation for each year is considered in conjunction with the means, a great deal of overlap is found to exist between individuals from each of the two companies, and access to the higher salary brackets was obtained by individuals with widely varying salary histories.

Figure 1 also indicates that there is a very close correspondence between the relative magnitudes of $\bar{k}_{..}$ and $\sigma_{k..}$ throughout the period studied. This high correlation is also characteristic of the pairs of $\bar{k}_{.c}$ and $\sigma_{k.c}$ for each company. Table 2 contains the Pearson product-moment correlation coefficients calculated between the pairs of means and standard deviations for the combined sample and each of the individual companies. These very high correlations imply that as the average salary increased, so did the dispersion of those salaries.

Another indication of the high degree of correlation between $\bar{k}_{..}$ and $\sigma_{k..}$ appears in Figure 11, which is a plot of the coefficient of variation for the combined samples $(\bar{k}_{..}/\sigma_{k..})$ as a function of calendar years. Figure 12 contains similar plots of the coefficient of variation $(\bar{k}_{.c}/\sigma_{k.c})$ for the individual companies. Whereas large differences in magnitude of this index existed for the data from individual companies, the range of values of the coefficient was quite restricted for the combined samples. The range of this index for the combined sample was 0.76, with a high value of 2.79 and a low value of 2.03. The average value of the coefficient for the period between 1938 and 1962 was 2.42.

The disparity in the shapes of the individual curves and the curve based upon the combined sample can be explained by examining each of the components in the coefficient. The numerator, $\bar{k}_{..}$, represents the average of a group of means, and hence its magnitude is bounded by the values of the separate $\bar{k}_{.c}$'s. The variation of the combined samples, $\sigma_{k..}$, represents the total or summed variation across all com-

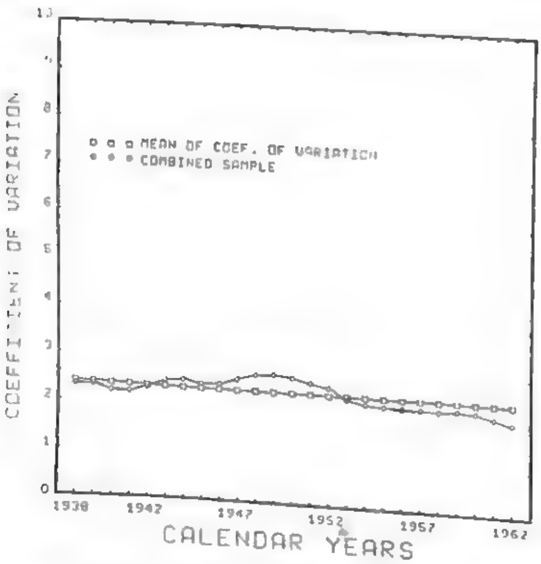


FIG. 11. Coefficient of variation $(\bar{k}/\sigma_k)$ for combined sample.

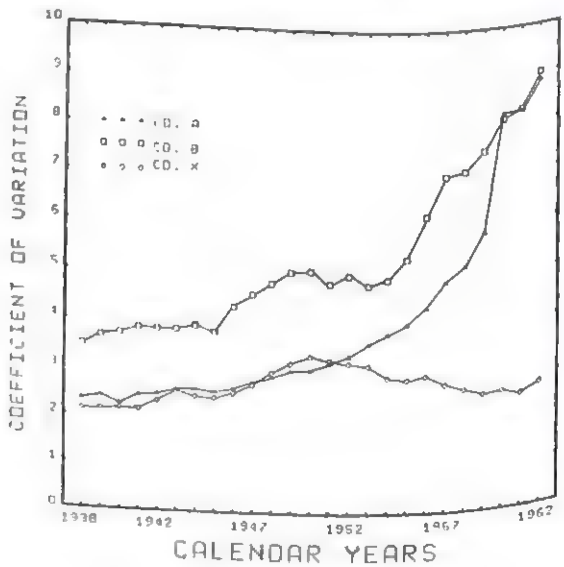


FIG. 12. Coefficient of variation $(\bar{k}/\sigma_k)$ by companies.

panies, and was larger than any of the individual company variations, $\sigma_{k.c}$. The increase in the denominator was responsible for the reduction in value of the coefficient calculated on the combined samples.

Table 2 also contains the correlation between the pairs of means and standard deviations of the increments for both the combined and the individual companies. Again, high correlations were found to exist for all pairs of $\bar{i}_{.c}$ and $\sigma_{i.c}$.

These results indicate that as the average salary and average increment increased over the years, so did the spread in salaries and increments, respectively. A high degree of correspondence was discovered between the relative magnitudes of $\bar{k}_{..}$ and $\sigma_{k..}$; the dispersion of salaries remained between one-half and one-third of the average salary.

Discussion

A variety of specific things stand out in these data. The first of them is the remarkably smooth, positively accelerated curve of total pay, and its consistency from sample to sample. This means that the executive can expect, on the average, larger earnings each year. From the point of view of the company, it could mean two related things: First, a cautious reluctance to pay much in the early years until a man has shown his worth. Second, to the extent to which salary measures the contribution to the company, the assumption that the contribution increases geometrically. It is also an expression of a kind of Weber-Fechner assumption that the raise must be a constant function of the absolute level. In many cases this has probably become the routine practice, rather than an expression of the more explicit elaborate rationale. In the extreme, it is hard to see the shapes of these curves as accurately reflecting contribution. Salary curves within a company generally build up smoothly to the chief executives. Yet the tremendous variation in the salaries of chief executives, published under Securities and Exchange Commission regulations, defies an easy explanation in terms of variable contribution either within an industry or across industries.

As far as these data show, the steady increase in pay displayed here may only be the factor which held these *S*s in the company long enough for them to be included in this study. Those who left, conceivably could have had quite different curves. Other data on pay make this unlikely.

A variety of curves of average pay over time is possible. A company might choose to keep the average increment roughly constant and the average pay curve linear. This would have different assumptions about the labor market and worth to the company, and about the motivation of employees. Or it might operate a declining average raise each year, hoping to hold individuals in early years with large raises. The choice of a rate of change in the curve has important motivational consequences, even though total cumulative career pay is the same. One might notice that the curves shown here do not fit the individual's demands very well. That is, pay is low when Junior's teeth are being fixed, sister is going to college, and the mortgage payments are due. After these crises have passed, pay increases.

It is also clear in these data that the variance of pay tends to increase as the average increases. This finding is in agreement with Brenner and Lockwood (1965). One startling result of this is that executives' salaries become more dissimilar over time, and, hence, an individual's position becomes more and more fixed relative to his fellows. If σ increases and N stays constant or shrinks, it takes a larger and larger increase in an individual's pay to change his rank order. This phenomenon may have important implications for individuals' needs for prestige and recognition.

It is also important to notice that the variance of pay is a function of two things: (*a*) the variance of increments, and (*b*) the correlation between raises and salaries. The variance of pay at Year *n* represents the variance of starting pay plus some fraction of the cumulative variance of the increments. The size of this fraction is a function of the nature of r_{ki}. The formula for the variance of a composite ($\sigma_{k+i} = \sigma_k^2 + \sigma_i^2 + 2\sigma_k\sigma_i r_{ki}$) makes it clear that the increase in the variance of pay at Year *n* is smallest when r_{ki} has been negative, and largest as r_{ki} approaches $+1.00$. We have probably underestimated the psychological possibilities in the cumulative character of pay.

It becomes clear that a variety of psychologically meaningful parameters of pay—the size of $\bar{k}$ at a given time, the shape of the curve

of $\bar{k}$, the size of σ_k, the size of σ_i at a given time and its rate of change—can be varied by a company without increasing the total salary expense over time. Yet the motivational leverage of alternative forms at constant cost has had remarkably little empirical investigation.

The measures of skewness and kurtosis in these data tell us more about how to handle the data themselves than they do about general characteristics of pay. Two of the samples (A and B) are from two groups of levels in a company. Under these circumstances it is hard to generalize the meaning of distributional shapes. The third (X) is, roughly, a random sample of a second company. Can we generalize? Both companies are superficially typical of industry in general in pay policies and administration. In this sense they are representative. However, the crucial question is whether they are representative on the more detailed internal parameters of pay—σ_i, r_{ki}, the rate of change of σ_i, and the like. Since there are no data on population characteristics for these variables, it is impossible to answer this question. It is hoped that subsequent research will report on these important parameters of pay so that psychological constructs in this area may be built on sound empirical bases.

2. The Correlation of Pay with Pay over Time

Among the things seen in the previous chapter was a marked progressive increase, over time, in the dispersion of salaries. An individual's salary, in absolute terms, becomes more and more unlike that of his fellow executives as length of service increases. An important question remains: as this dissimilarity in absolute salaries increases, what becomes of the relative position of an individual in the array of salaries? Does pay predict pay?

A number of psychological questions are implicit in this relationship of pay with pay. (*a*) For the individual it is the question of "who gets the raises?" since pay on Year 2 is equal to pay on Year 1 plus the raise from 1 to 2. For the individual the raise is the reward of merit. Raises partake of the whole field of variable reinforcement in psychology, and much of the incentive character of pay must reside here. Further, as raises accumulate, they idctate relative position in the array of salaries, with its attendant implications for need satisfaction in the form of prestige, recognition, and achievement. (*b*) From the side of the company, the relation between raises and past pay uncovers some implicit assumptions about the consistency of good or useful performance. If raises are held to be related to performance—a position that is often, though not invariably, taken—the relation of pay with pay becomes a measure of consistency of performance, and makes contact with a long tradition of research and theory in differential psychology. (*c*) The correlation of pay with pay can also be seen as a problem of prediction. If the relationship between pay and performance mentioned above is assumed, as Brenner and Lockwood (1965) do, pay can be seen as an intermediate criterion in the Thorndikean sense (Thorndike, 1949) as well as a predictor. (*d*) The shape of the curve of the correlation of a criterion pay with lagged years has interesting possibilities. Various studies (Fleishman & Fruchter, 1960) have suggested shifts in factor loadings in performance over time. In the present data, the executives move, over time, through a variety of levels. It has often been suggested that different organizational levels require quite different skills. If this is so, and if pay represents performance, the curve of prediction of pay by past pay may show discontinuities or inflections as a function of these changing factors associated with organizational structure. (*e*) Finally, as has been pointed out before, the composite character of pay is again important. Pay this year is last year's pay plus a raise. Consequently the correlation of pay with pay is appropriately represented by the formula for composites,

$$r_{k(k+i)} = \frac{\sigma_k + \sigma_i r_{ki}}{\sqrt{\sigma_k^2 + \sigma_i^2 + 2\sigma_k\sigma_i r_{ki}}},$$

where k is the pay on the previous year and i the increment. (The psychological meaning of these parameters is briefly discussed in the previous chapter.) A glance at the formula makes it clear that when $r_{ki} = 0$, the correlation between successive years is determined by the ratio of the two variations, that is, when $r_{ki} = 0$,

$$r_{k(k+i)} = \frac{\sigma_k}{\sqrt{\sigma_k^2 + \sigma_i^2}}.$$

Further, if $r_{ki} = 1.0$, $r_{k(k+i)} = 1.0$. In terms of salary administration these two policies mean that raises may either be distributed at random with respect to past salaries or that those with the highest salaries get the biggest raises. Obviously, all the other policies between $r_{ki} = -1.0$ and $r_{ki} = +1.0$ are possible. This dissection of the formula for the correlation of composites emphasizes the role of r_{ki}—the rules for awarding raises to employees with different salaries—as a major factor in the control of the correlation between salaries on different years. In Chapter 1, the role of this same r_{ki} in the variance of composites was stressed. Among the measures normally monitored in wage and salary administration, this vital variable has been largely overlooked.

In the only other longitudinal study of pay that comes immediately to hand, Brenner and Lockwood (1965) suggested that pay becomes a better predictor of pay as length of service increases. In their report on 52 aircraft engineers over 20 yr. they showed an increase in the correlation between adjacent years as tenure increased. This pattern might be interpreted on the basis of (*a*) consistency of performance plus (*b*) an increasing awareness of the relative value of each individual. The empirical fact of the relationship can be examined in the data presented here.

The matrix of intercorrelations of each year with each other year was calculated for each

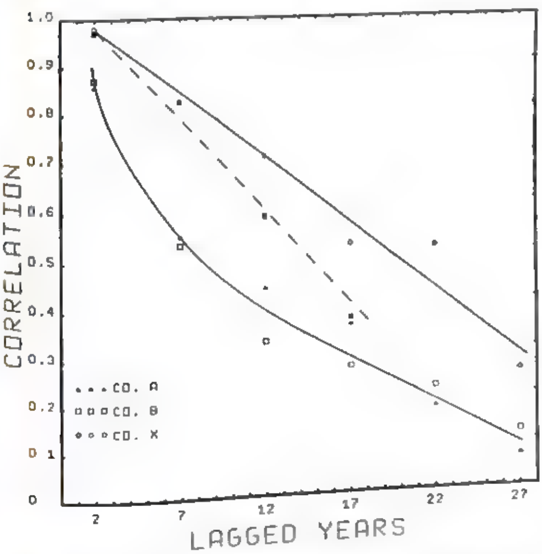


Fig. 13. The correlation of pay with pay over lagged years. (Data from Brenner and Lockwood, 1965, are plotted with asterisks for comparison.)

of the three groups of managers previously studied. Figure 13 shows the correlation of a final level of pay—taken as a criterion—with pay on earlier years. The pay on years 1956–1962 was chosen as the criterion. Salaries on these years were correlated with the pay 2 years before (i.e., 1954–1959), 7 years, 12, etc. Each point on the graph represents the average of the correlations for each year's salary in the criterion period with the salary x years previous to it (e.g., the correlation for 2 lagged years represents the average of the correlations between 1961 and 1959, 1960 and 1958, 1959 and 1957, . . ., and 1956 and 1954). This average was calculated by converting all the intercorrelations to z scores utilizing Fisher's transformation, averaging the z scores, and converting the average z to a figure for a product-moment correlation coefficient.[5]

It is immediately evident that there is a steady decrease in the predictive power of past pay in all three companies. This decline in correlation implies that the executives studied have changed their relative positions with regard to the amount of pay they receive, and that the amount of change which occurred increased over time. Company X seemed to be the least affected. The correlations are uniformly higher throughout the 27 lagged years. The relative pay positions in this company seem to have remained the most stable. Company A and Company B show very similar

[5] The use of the Pearson product-moment correlations was deemed permissible despite the high degree of skewness present in the yearly distributions of salary because the direction and magnitude of skewness was fairly similar across the years (see first chapter for a report of the index of skewness). Ghiselli (1964) points out that the maximum range of values of the Pearson r will be affected by departures from normality only if the distributions have different shapes (i.e., if there are large differences in the skewness of the distributions of the variables being correlated). In Company A the range of the index of skewness was 1.24, with all but one distribution being positively skewed. In Company B, 28% of the distributions were negatively skewed. The rest of the distributions had index values greater than 0.0, and the range of values for skew was 1.56. And in Company X all distributions were positively skewed. The range of the index was 1.66. It will also be noted that the lagged years run to 27. This final point runs beyond the requirement of 40 Ss in each group. This is the only computation utilizing these data. It is included here because the 27-yr. point is a smooth continuation of the line generated by the earlier points.

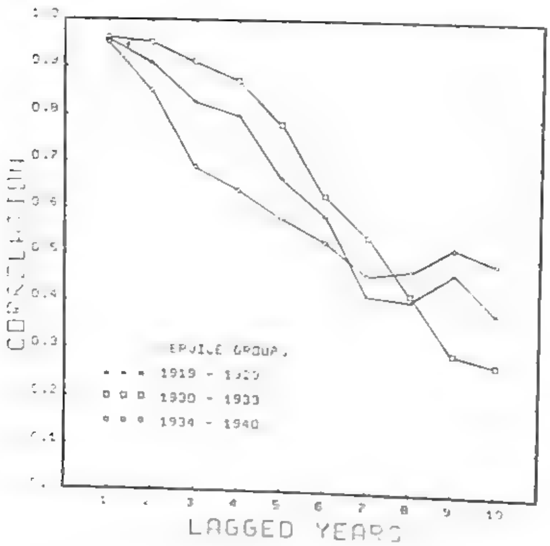


Fig. 14. Correlation of pay with pay for three different service groups for lagged years in the period 1958–1949.

sharp decreases in correlation for the first 7 lagged years. The rate of decline stabilizes thereafter, so that the slopes of the curves for all three companies are somewhat similar.

For the purpose of comparison, the data presented by Brenner and Lockwood (1965) are plotted in Figure 13. Their published matrix of intercorrelations was handled in a manner identical to the data reported here. Only 17 yr. of lag are available in the Brenner and Lockwood data.

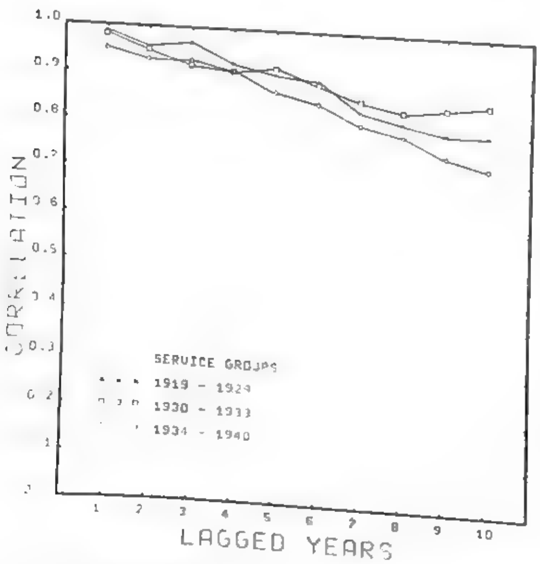


Fig. 15. Correlation of pay with pay for three different service groups for lagged years in the period 1951–1942.

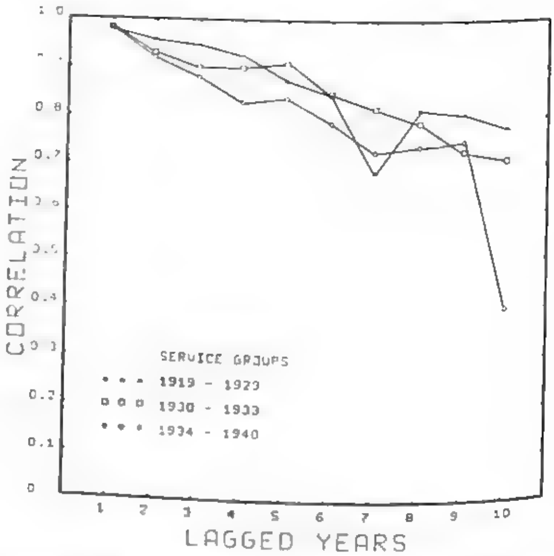


Fig. 16. Correlation of pay with pay for three different service groups for lagged years in the period 1945–1936.

The decline in correlation of pay with pay may be a function of particular stages in a man's career—or, in other words, of length of tenure. To test this hypothesis, the Company A sample was divided into three groups having different dates of employment. The groups were hired between 1919 and 1929, 1930 and 1933, and 1934 and 1940. The three periods were chosen arbitrarily to give roughly equal Ns and a wide chronological span. Inter-r's were computed separately for each group, and the results appear in Figures 14 through 16.

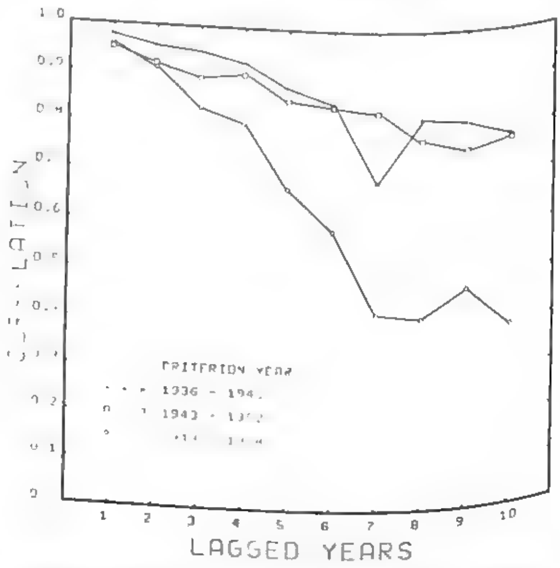


Fig. 17. Correlation of pay with pay in adjacent years from 1962–1952.

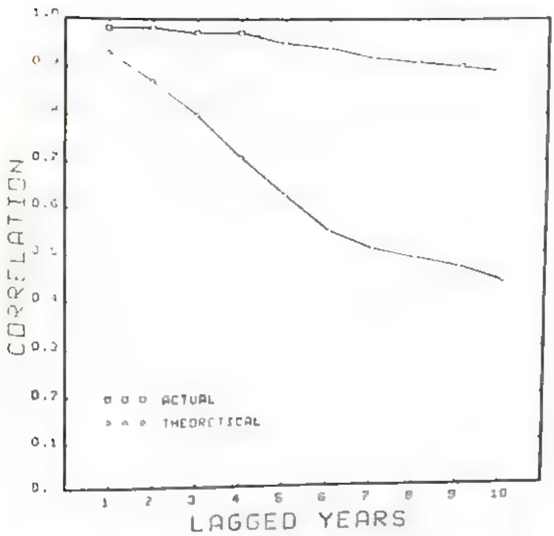


Fig. 18. Actual correlation of pay with pay over lagged years compared to an assumption of random raises. (Company A, 1951–1941)

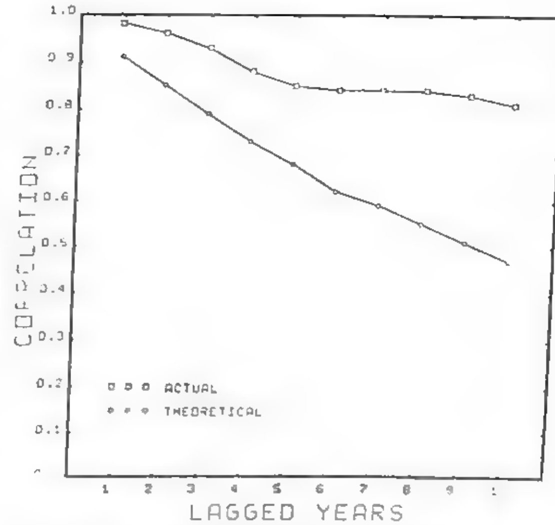


Fig. 20. Actual correlation of pay with pay over lagged years compared to an assumption of random raises. (Company B, 1951–1941)

These figures show the $r_{k(k+i)}$ for lagged years for three selected chronological periods. The plots for all three service groups in a given chronological period are very similar. The three periods are rather different from one another, but within each the three service groups are remarkably similar. There is no apparent effect of tenure on the correlation of pay with pay. Brenner and Lockwood compared the correlations of adjacent years as tenure increased and found a positive relationship. This would lead one to expect the 1919–1929 group to have the highest r's in a given period, the 1930–1933 group next, and so on. Figure 17 presents the data for each company in the same form as the Brenner and Lockwood data. It is clear that the companies sampled here do not show the phenomenon.

What causes the slope of the curve of $r_{k(k+i)}$ over time? What causes the decline in the predictability of pay by pay? The correlation $k(k+i)$ is importantly influenced by r_{ki}. To

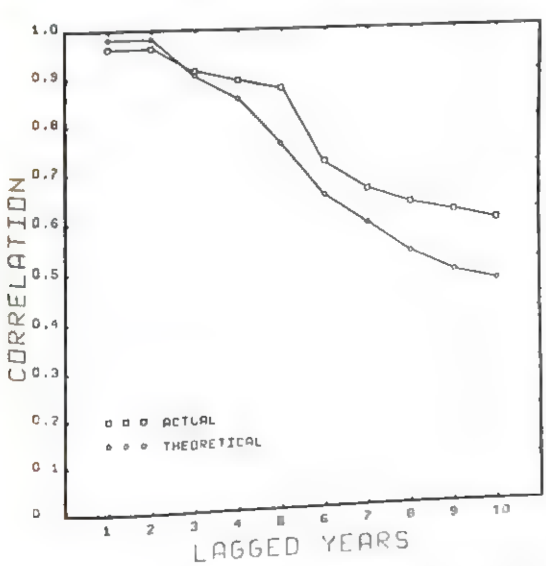


Fig. 19. Actual correlation of pay with pay over lagged years compared to an assumption of random raises. (Company A, 1958–1948)

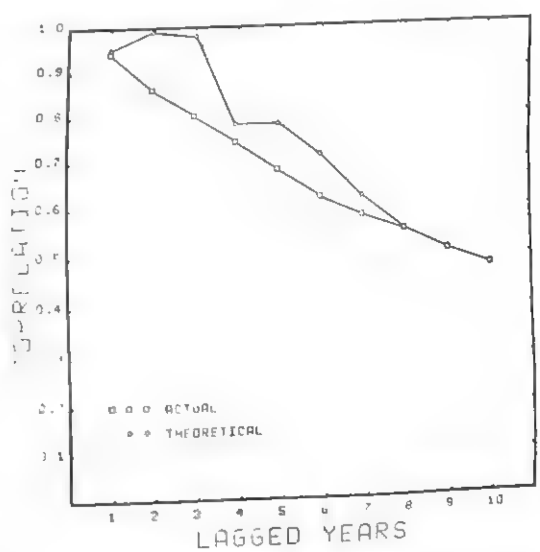


Fig. 21. Actual correlation of pay with pay over lagged years compared to an assumption of random raises. (Company B, 1958–1948)

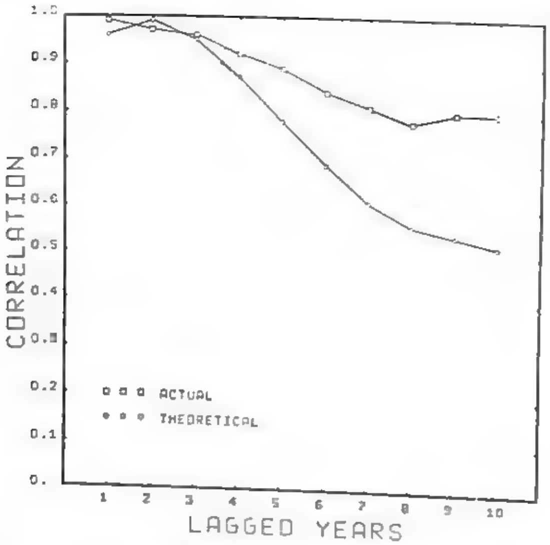


FIG. 22. Actual correlation of pay with pay over lagged years compared to an assumption of random raises. (Company X, 1951–1941)

examine the effect of this correlation on the correlation of salaries for different years, r_{ki} was assumed to be zero, and a series of theoretical curves was computed from the ratio of the variations $(r_{k(k+i)} = \sigma_k/\sqrt{\sigma_k^2 + \sigma_i^2})$ for lagged years. Figures 18 through 23 show the relationship between these theoretical curves assuming a random distribution of raises and the observed correlations. The theoretical curves provide a good fit for Companies A and B subsequent to the mid-50s; before that they

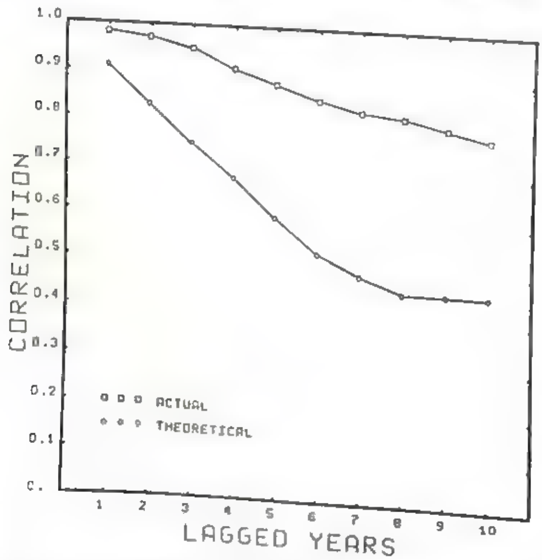


FIG. 23. Actual correlation of pay with pay over lagged years compared to an assumption of random raises. (Company X, 1958–1948)

do not. Company X data did not fit the theoretical assumption consistently for any of the periods. In 1953, Companies A and B introduced a change in performance evaluation which was probably responsible for this change.

Prior to 1953 the appraisal was, roughly, of the "is he a good man?" type. After 1953 a very serious effort was made to evaluate the contribution of the man on the job. Since the potential contribution of jobs and the difficulty of accomplishing them varied widely, and since assignment to jobs had a random character due to the historical accidents of need and availability, a random element was introduced into appraisal and hence into pay.

The effect of the change in policy is clearly seen in Figure 24, which shows the actual r_{ki} for all pairs of successive years between 1948 and 1958 and for each of the three companies. The r_{ki} pre- and post-1953 for each of the three companies appear in Table 3. Company X remains fairly constant on the average. Companies A and B change markedly. The considerable instability in r_{ki} overall might also be noted.

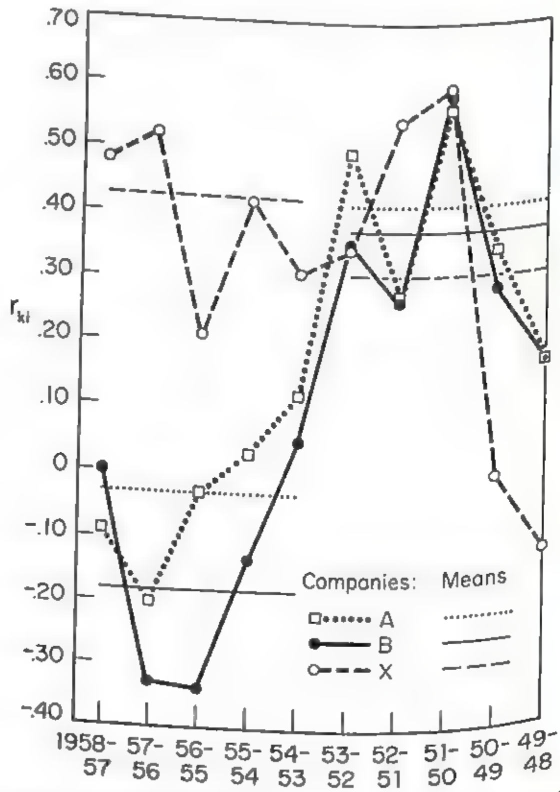


FIG. 24. The correlation of pay with raises on pairs of successive years for three companies from 1948–1958.

TABLE 3

Average Correlation of Salaries and Raises by Company

	$\bar{r}_{ki}$	
	1948–1953	1954–1958
Company A	.42	−.06
Company B	.38	−.18
Company X	.32	.42

An attempt was made to discover whether any clusters of career salary patterns existed by factor analyzing the intercorrelations among salaries over the years. The purpose of this analysis was to determine whether the correlations between pay distinguished different groups of individuals. The procedure utilized was analogous to what Guilford (1954) has described as the *Q* technique of factor analysis. The results of this procedure indicated that greater than 95% of the variance could be accounted for by a single factor. This factor represented the salary previously paid to the employee. That is, groups of executives could not be differentiated because each individual's salary was composed of the same two components: past pay and an increment. Other attempts were made to discover additional factors by alternative procedures, but so far these attempts have been unsuccessful.

The correlation of pay with pay declines regularly and fairly rapidly. Companies A and B behaved similarly, with a point of inflection probably associated with the shift in performance evaluation policy mentioned previously. Company X and the Brenner and Lockwood data show similar declines probably best described by straight lines. The consistency within each of the two pairs of data over the same period strongly suggests that $r_{k(k+i)}$ is a function of company policy and not of the exigencies of historical periods. Several things flow from this decline in $r_{k(k+i)}$:

(*a*) The decline is smooth. Except for the curve in Companies A and B there are no serious breaks. This does not suggest—though it does not rule out—shifting factor loadings from management level to management level. The correlation can also be thought of as the ratio of common factors in the job in Years k and $k+1$ to the total number of factors in the 2 yr. This view, too, leads to the suggestion of a relatively smooth change in job factors from year to year.

(*b*) It should be pointed out that these data are not ideal for this analysis of factors. The pay confuses changes in level and penetration of the salary range for the level. Some of the smoothness may be an artifactual result of the confounding of these two sources of increases.

(*c*) The sharp decline in correlation raises questions about the qualities for which employees are paid. We generally expect a fair correlation in the performance of mature individuals over time. The decline in correlation here suggests either that this accustomed consistency of performance is absent or that the pay or raises (or both) are unrelated to performance. If the latter is true it has real implications for the view of pay as a reward in the traditional model of behavior change.

(*d*) While the correlation between pay and pay declines, it is still substantial. From the side of the company, it provides a real implement for career planning and development. From the point of view of the individual, the substantial correlation means a stable relative position in the arrays, with its attendant motivational implications.

The approach of r_{ki} to 0 has disturbing psychological implications. Baldly, it means that raises are randomly distributed with respect to performance. It echoes the lack of consistency in (rewarded) performance just mentioned. The awareness that having done well is no better predictor of the future than having done poorly must take most of the incentive out of the pay. Even when, as has been suggested in Companies A and B, the pay is given not for individual performance but for performance on the particular job with a random assignment to jobs, the heart of the incentive is gone. It takes only a little algebraic manipulation to see that if $r_{ki} = 0$, $r_{ii} = 0$. On the job it must take relatively little experience with the system to realize the same thing.

Psychologists are seldom in a position to say how data *ought* to be in the real world. Here it seems possible. The correlation *ki ought* to be positive and significantly different from zero. Two lines of reasoning support the point

of view: (*a*) It would better represent a consistency of performance which probably exists. (*b*) It would rescue the individual from the morass of insecurity engendered by the complete lack of prediction of raises by raises.

Psychologists must consider the relation between the value of r_{ki} and techniques of performance evaluation. The old (bad) model was to give a man a raise because you liked his looks. The next year his looks had not changed and he got another. The correlation i, i equaled +1.0, and r_{ki} approached it. As more and more factors are introduced into performance appraisal, the denominator of the formula for correlation stated in terms of common factors grows, without any necessity that the numerator keeps pace. Only a most vigilant management of the appraisal system will keep it from vitiating the incentive in the pay policy.

Brenner and Lockwood suggested and demonstrated a positive relationship between tenure and the correlation between adjacent years. Indeed, it is a reasonable hypothesis. Late pay may be thought of as representing seasoned and selected performance, while early pay has all the vagaries of the untried and inexperienced. All that can be said is that these data do not support the hypothesis. The correlation and its changing values seem to be more a function of particular company policies than of the development of individuals.

Examining the correlation of pay with pay highlights the psychological leverage available to salary administrators in little-noticed parameters of pay, r_{ki} and σ_i. σ_k is an extant quantity at any given time (the dispersion of last year's salaries). Consequently, control over r_{ki} and hence over $r_{k(k+i)}$ is a function of σ_i. Attention has been given to σ_i before in terms of the motivational leverage inherent in policies about raises. Patton (1961) has emphasized the need for recognizing and rewarding outstanding performance with raises which are directly related to the executives' contributions for the year. Patton's thesis is intimately connected with the motivational aspects of σ_i. However, little is said about the effect of σ_i on relative pay over time. It appears to be a crucial motivational component, and to provide a handle for the management of relative pay positions. It is surely not unimportant to point out that this motivational objective of the company can be achieved without suggesting a change in $\bar{k}$ or $\bar{\imath}$, the salary bill.

3. A Statistical Model of Certain Parameters of Pay

We have seen the behavior of certain parameters of pay over time, and the performance of pay viewed as a composite variable, that is, the fact that pay on any given year is equivalent to pay on the previous year plus or minus some increment ($k_n = k_{n-1+i}$). Based upon this premise one is able to restate the formulae for composites so that the standard deviation of pay on succeeding years is

$$\sigma_{k+i} = \sqrt{\sigma_k^2 + \sigma_i^2 + 2\sigma_k\sigma_i r_{ki}},$$

and the correlation of pay on successive years is

$$r_{k(k+i)} = \frac{\sigma_k + \sigma_i r_{ki}}{\sigma_{k+i}}.$$

The merit of this conceptualization is that the parameters of pay are united parsimoniously in a statistical model of composite variables, thereby formalizing the relationships which exist among the parameters. Presumably, at least two areas of compensation research should be enhanced with the adoption of this statistical formulation of pay. First, rather than treating pay holistically (macroscopically), future research on compensation may be designed to determine which of the many parameters of pay is responsible for its relation to other variables. For example, a program of research structured in terms of the statistical model could facilitate the discovery of the linkage between various characteristics of pay and its incentive value. Hence, the model provides the basic constructs for a more microscopic analysis of pay.

Second, this mathematical formulation facilitates operationalism in the realm of compensation practices. Different compensation policies may be defined in terms of the manipulation of specific parameters. As a consequence of these refinements, the behavior of pay may be predicted as a function of different remunerative plans. This chapter is devoted to exploring the usefulness of the model in this second area of defining and predicting the effects of different compensation policies.

The mathematical definition and the psychological meaning of each of the parameters identified in the above formulas have already been discussed in Chapters 1 and 2. The variables involved can be divided into several types for the present analysis:

1. Some are measures of extant quantities at any given time, and hence are unavailable for manipulation. Examples are $\bar{k}$ and σ_k, the mean and dispersion of last year's salary. They are largely disregarded here.

2. A second group is made up of the variables which the company manipulates as it tries to make its compensation do whatever it is that is aimed at. For instance, σ_i and $r_{k,i}$, the variation in raises and in who gets them, stand out in this class. Since they are manipulable in this sense, they will be treated here as if they were independent variables.

3. Another group is made up of composites of the first two: σ_{k+i} and $r_{k,k+i}$, for instance, measure the effect on last year's salaries of the way in which this year's raises were awarded. They are treated here as dependent variables for two reasons. First, they are the direct result of manipulation of the independent variables. Second, they bear much of the psychological impact of pay. σ_{k+i} brackets the scope of an individual's salary expectations; $r_{k,k+i}$ contains the probability that he will change his relative salary status over adjacent fiscal periods.

It is worth noticing that $\bar{i}$ does not appear as a variable on either side of the equation. This is important, since Σi represents the increase in the wage bill. Above some discriminable minimum, Σi, the cost of raises is not an important variable in determining the way the things management manipulates (σ_i, $r_{k,i}$) affect the psychological variables (σ_{k+i}, $r_{k,k+i}$). $\bar{i}$, of course, has a real influence on $\bar{k}$, since pay over the years is simply the starting salary plus the accumulation of increments.

Let us look at some compensation policies in these terms. By "compensation policy" we mean the company's active manipulation of pay to achieve specified quantitative relationships among salaries (and hence psychological effects on the modification of behavior). We will examine some policies with respect to the two independent variables σ_i and r_{ki} on the dependent variables σ_{k+i} and $r_{k,k+i}$, and consider the psychological meaning of the resultants.

Policies Affecting σ_i

S1. $\sigma_{i_1} = \sigma_{i_2} = \sigma_{i_3} = \cdots = \sigma_{i_n}$. This policy maintains equal standard deviations of the increments each year. The same individual may not receive exactly the same increase each year: this is a function r_{ii}, not σ_i. The amount of an individual raise is not strictly determined by this policy. The employee's aspiration as to the absolute amount of salary he can earn is still nearly unlimited. However, the relative amount he can be awarded is fixed in relation to the raises given to fellow employees regardless of the distributions of relative contributions and development of individuals from year to year. The effect on incentive would seem to be a function of the absolute size of σ_i selected; the larger the value of σ_i, the greater the ability of the company to reward its employees differentially.

S2. $\sigma_{i_1} < \sigma_{i_2} < \sigma_{i_3} < \cdots < \sigma_{i_n}$. This policy indicates that the magnitude of σ_i increases each year. It is likely to occur when each individual is awarded an increment proportional to present salary. That is, a linear transformation of the salaries produces a nonlinear (logarithmic) transformation of the values of σ_i. Such a policy might also be designed to recognize that the relative contributions of some individuals are increasing while those of other employees are decreasing or remaining the same over time. These conditions would increase the value of σ_i each year assuming that raises were positively correlated with performance.

S3. $\sigma_{i_1} > \sigma_{i_2} > \sigma_{i_3} > \cdots > \sigma_{i_n}$. This policy is designed to steadily reduce the dispersion of increments each year. This type of policy would probably be reserved for a situation in which there was a definite ceiling on performance so that with time each person's contribution becomes more like that of others. This policy is not totally unrealistic; the requirements of some standardized jobs demand that an attempt be made to equalize the wages of all workers performing the job. However, this policy probably does not represent common practice, especially among salaried employees.

Policies Affecting $r_{k.i}$

R1. $r_{ki} = a$, where $0 < a \leq 1.00$. That is, increases in pay bear a constant positive relationship to pay on the previous year. This type of policy is both familiar and very reasonable. First, it represents the fact that, in general, relative performance is positively correlated over time. This is usually reflected in the stability of relative pay over the years. Second, many organizations award raises which are approximately constant fractions or percentages of current salary. In this situation, the value of r_{ki} approaches 1.00, in which case $r_{k(k+i)}$ also approaches a value of 1.00. Note that the higher the value of r_{ki}, the more stability present in relative pay over time. This stability would appear to have clear motivational implications. Regardless of any effort one makes to change his relative monetary status by means of spurts in performance, pay position will remain relatively fixed over time whenever the value of r_{ki} approaches 1.00.

R2. $r_{ki} = 0.0$. That is, increases in pay are randomly assigned with respect to pay on the previous year. This type of policy ignores any persistent trends in relative performance over time and could stem from several circumstances: (*a*) the standards used by management as a basis for awarding raises vary from year to year (see Chapter 2); (*b*) the individual moves from one job to another almost every year, each job having different requirements. The individual's talents for one job may be quite inappropriate for the tasks of another and so he would perform well on one job and poorly on the other; (*c*) performance is measured without any reliability at all; therefore, evaluation of an individual's performance varies from one year to the next.

R3. $r_{ki} = b$, where $-1.00 \leq b < 0.0$. That is, increases in pay bear a constant negative relationship to pay on the previous year. This policy would most likely be employed in a conscious attempt to equalize pay or eliminate existing inequities in salary. Salary differences would be reduced because this policy dictates that the largest raises be awarded to those individuals with the smallest current salary, whereas the individuals who currently receive the highest salary would be awarded the smallest raises. This may also occur where some people have reached the upper portions of their salary brackets. The amount of recognition that can be awarded in the form of raises is small. Any major recognition of their performance will most likely be in the form of a promotion.

In this manner, we can examine different compensation policies in terms of a statistical model of pay. Certain remunerative policies will be investigated by examining the configuration of independently manipulable parameters required to maintain such policies and the effects of these policies on both the dispersion and relative stability of salaries over time. The object was to obtain an understanding of the long-range effects produced by specific modifications of the parameters of pay and was not concerned with the particular history of compensation practices of any given (existing) company.

In order to do this, a group of computer programs were developed which incorporated the statistical formulas which constitute the model of pay as well as a series of instructions for systematically modifying the parameters of pay in order to simulate a variety of compensation policies. Simulation of the different compensation policies was accomplished by varying the values of σ_i and r_{ki}. Policy S1 was simulated by keeping the value of σ_i constant, that is, equivalent to initial magnitude, throughout the calculations. Policies S2 and S3 were simulated by successively increasing or decreasing, respectively, the value of σ_i by one of four constant percentages: 5, 10, 15, or 20%. The three policies concerned with the magnitude of r_{ki} were simulated by setting r_{ki} equal to values which varied from -1.00 to $+1.00$ in steps of 0.25.

Calculations were made possible by supplying the computer with initial values for all the parameters contained within the model. σ_k was initially given a value of 1.00. On each successive calculation, the value of σ_{k+i} was substituted for σ_k. That is, σ_k for any year was set equal to the value of σ_{k+i} calculated on the basis of data from the previous year.

Four initial values of σ_i were selected so that the ratio σ_i/σ_k initially represented a full range of meaningful relations between these two parameters. σ_i/σ_k was assigned values of either

.1, .4, .7, or 1.0 at the beginning of the computations.

The effects of these various compensation programs were observed on two of the parameters of pay, σ_{k+i} and $r_{k(k+i)}$, for successive periods in a 25-period span. It may be convenient to think of the periods as years, but it is, of course, arbitrary.

The results of each policy regarding σ_i were observed in combination with each of the policies regarding the value of r_{ki} and each of the initial ratios of σ_i/σ_k. Thus, the total configuration was analogous to a completely crossed analysis-of-variance design with three factors: σ_i, r_{ki}, and σ_i/σ_k. The design can be described as a $9 \times 9 \times 4$ design matrix and may be graphically represented as in Figure 25.

For each cell of the matrix depicted in Figure 25 we have obtained 25 pairs of values of σ_{k+i} and $r_{k(k+i)}$. σ_{k+i} and $r_{k(k+i)}$ have been plotted as a function of time for all 324 combinations of treatments. Space limitations prevent presentation of the graphs of the results of all treatment combinations. However, Figures 26 through 31 sample the design matrix and illustrate the major findings on the behavior of the parameters σ_{k+i} and $r_{k(k+i)}$. Observe that for each combination of policies regarding r_{ki} and σ_i there are four curves reflecting the effect of σ_i/σ_k which demonstrate the short-run phenomenon we have named convergence.[6] From the point of view of the organization, the phenomenon of convergence for the curves of $r_{k(k+i)}$ implies that the predictive power of pay on pay on the succeeding year will not vary markedly after x number of years.

This same point can be restated in terms of its meaning for the individual. When convergence is reached, the probability of change in relative salary position will not change markedly.

[6] Although it is possible to derive algebraic solutions for the asymptote of many of the functions of $r_{k(k,i)}$ and $\sigma_{k,i}$, and thereby demonstrate the similarity among the curves in terms of their behavior at t_{infinity}, these mathematical abstractions have little meaning because the results must reflect the fact that the worker is employed for a limited period of time; from 25 to 50 yr. at most. Hence, the term "convergence" has been selected to signify the apparent leveling off and converging of the curves over the short run.

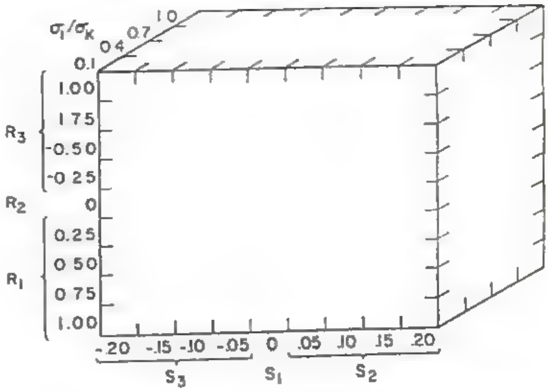


Fig. 25. Schematic diagram of the variables considered and the specific values employed for computation.

Effects of r_{ki}

The effects of r_{ki} upon $r_{k(k+i)}$ can be clearly differentiated into those for $r_{ki} \geq 0.0$ and those for $r_{ki} < 0.0$.

(*a*) When $r_{ki} > 0.0$, it can be demonstrated algebraically that $r_{k(k+i)}$ becomes asymptotic at a value of 1.00 as time goes to infinity. However, for the span of time investigated, the curves of $r_{k(k+i)}$ level off, or converge, at a value less than 1.00, but always positive, depending upon the values of r_{ki}. While the

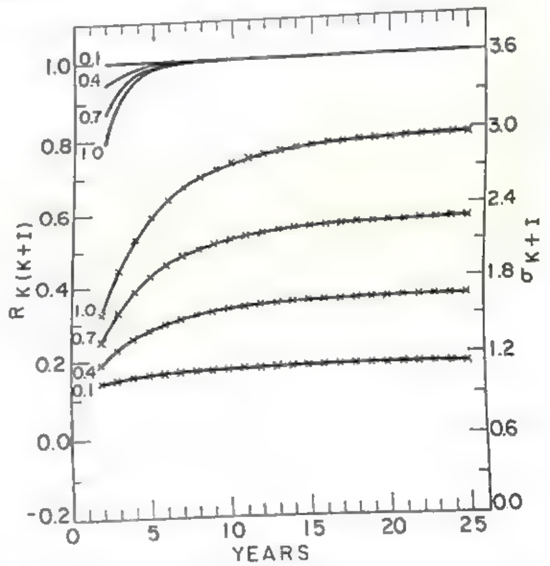


Fig. 26. Sample illustration of the variation of σ_{k+i} and $r_{k(k+i)}$ as a function of time for various policies where raises (r_{ki}) = .25 and variation in pay (σ_i) is progressively decreased by 20% each year. (A family of curves represents different initial ratios of σ_i/σ_k. The upper family of curves is the correlation of pay with pay. The lower family—hatched lines—is variation in pay.)

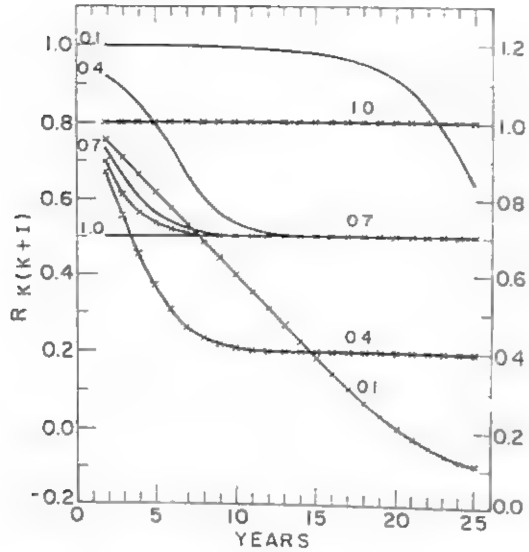


FIG. 27. Sample illustration of the variation of σ_{k+i} and $r_{k(k+i)}$ as a function of time for various policies where raises $(r_{ki}) = -.50$ and variation in pay (σ_i) is constant. (A family of curves represents different initial ratios of σ_i/σ_k. The upper family of curves is the correlation of pay with pay. The lower family—hatched lines—is variation in pay.)

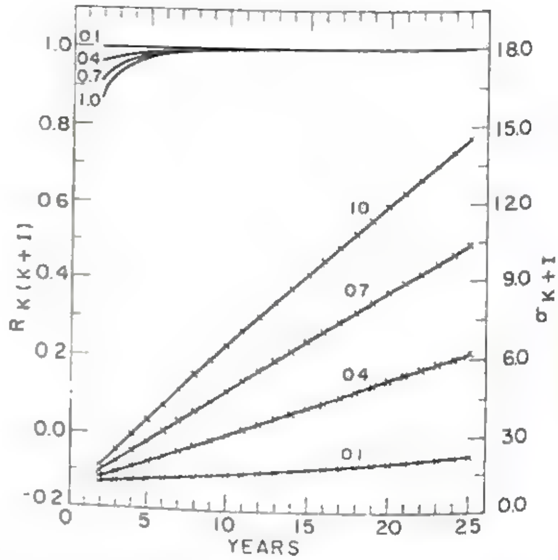


FIG. 29. Sample illustration of the variation of σ_{k+i} and $r_{k(k+i)}$ as a function of time for various policies where raises $(r_{ki}) = .50$ and variation in pay (σ_i) is constant. (A family of curves represents different initial ratios of σ_i/σ_k. The upper family of curves is the correlation of pay with pay. The lower family—hatched lines—is variation in pay.)

curves of $r_{k(k+i)}$ always converge, the time it takes to stabilize varies considerably as a function of r_{ki}. The higher the value of r_{ki}, the faster the approach to a stable $r_{k(k+i)}$. The limiting case in this argument occurs when $r_{ki} = 1.00$, under which circumstances $r_{k(k+i)}$ immediately assumes a value of 1.00.

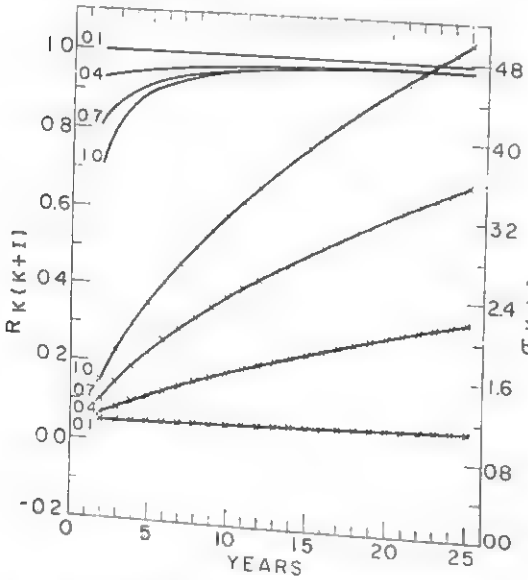


FIG. 28. Sample illustration of the variation of σ_{k+i} and $r_{k(k+i)}$ as a function of time for various policies where raises $(r_{ki}) = 0$ and variation in pay (σ_i) is constant. (A family of curves represents different initial ratios of σ_i/σ_k. The upper family of curves is the correlation of pay with pay. The lower family—hatched lines—is variation in pay.)

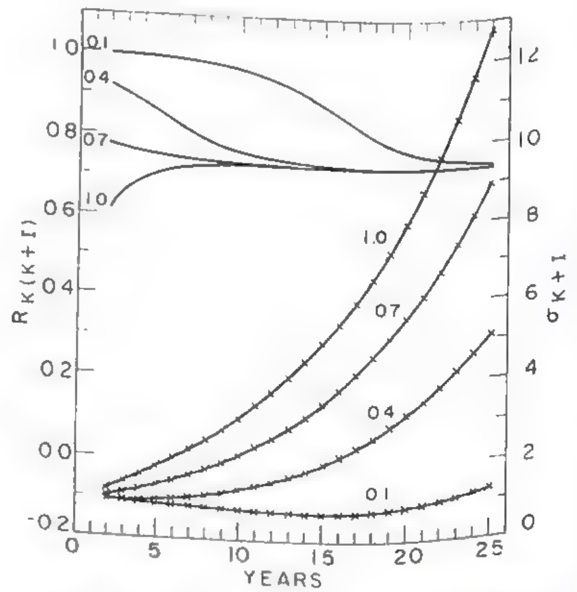


FIG. 30. Sample illustration of the variation of σ_{k+i} and $r_{k(k+i)}$ as a function of time for various policies where raises $(r_{ki}) = -.25$ and variation in pay (σ_i) is progressively increased by 10% each year. (A family of curves represents different initial ratios of σ_i/σ_k. The upper family of curves is the correlation of pay with pay. The lower family—hatched lines—is variation in pay.)

The fact that $r_{k(k+i)}$ has an asymptote of 1.00 under Policies R1 and R2 may be attributed to the steady growth of σ_{k+i}. The formula for $r_{k(k+i)}$ can be reduced to two parts: σ_k/σ_{k+i} and $\sigma_i r_{ki}/\sigma_{k+i}$. At infinity the latter part is essentially equivalent to 0.0 while the former is equivalent to 1.0. In common sense terms, when the dispersion between salaries gets very large, relative monetary positions will remain unchanged following the dispensation of raises due to the small size of σ_i in relation to σ_k. That is, even the largest raise will not be big enough to cause the interchange of salary position within the group.

Convergence at a level of $r_{k(k+i)}$ close to 1.00 affirms the inevitability that the relative monetary positions of the members of a group will become permanently fixed. This would seem to have profound motivational implications for the individual who must become resigned to the fact that he will always occupy the same position in terms of salary in relation to fellow employees, no matter how exceptional his performance.

(*b*) When $r_{ki} < 0.0$, the effects on $r_{k(k+i)}$ are heavily dependent upon the policy regarding the change in σ_i. The only general statement

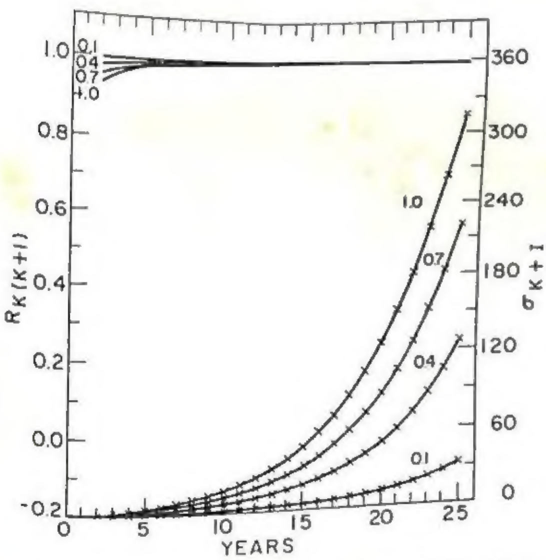


FIG. 31. Sample illustration of the variation of σ_{k+i} and $r_{k(k+i)}$ as a function of time for various policies where raises $(r_{ki}) = .75$ and variation in pay (σ_i) is progressively increased by 20% each year. (A family of curves represents different initial ratios of σ_i/σ_k. The upper family of curves is the correlation of pay with pay. The lower family—hatched lines—is variation in pay.)

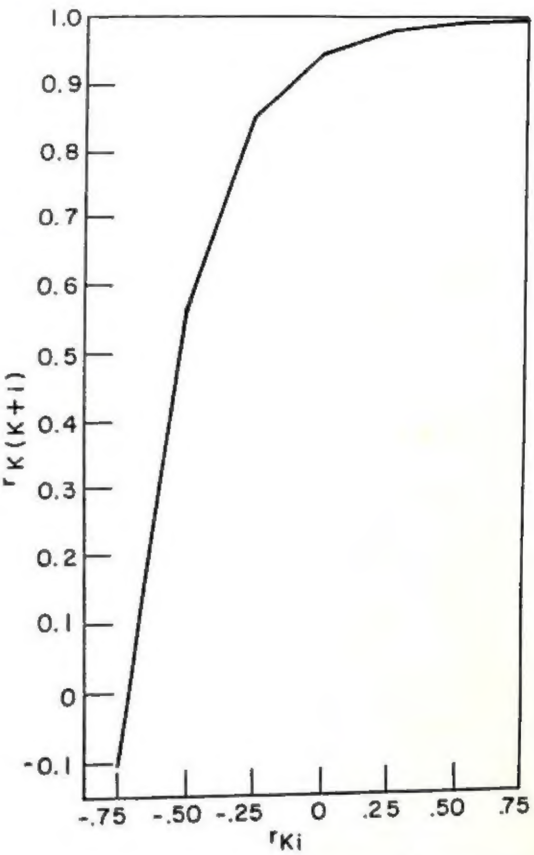


FIG. 32. The mean level of convergence of $r_{k(k+i)}$ averaged across values of σ_i plotted as a function of r_{ki}.

that can be made is that the value of $r_{k(k+i)}$ has fairly well stabilized after 25 yr.[7] The level at which convergence and relative stability occurs is, however, a function of the policy controlling the change in σ_i; that is, convergence occurs at a higher level of positive correlation for a given r_{ki} as the compensation policy is varied from S3 to S1 to S2.

Regardless of the policy controlling the change in σ_i over time, the greatest modification in the level of convergence of $r_{k(k+i)}$ for a given amount of change in r_{ki} occurs when the value of r_{ki} is negative. Figure 32 contains a plot of the magnitude of the level of convergence for $r_{k(k+i)}$ as a function of r_{ki}. Since the effects were similar across policies affecting σ_i, we have averaged the levels of convergence across Policies S1, S2, and S3.

If it is assumed that recognition of a change in the stability of relative salaries is directly

[7] For example, the asymptote for $r_{k(k,i)}$ when $r_{ki} < 0.0$ and Policy S1 prevails is: $r_{k(k,i)} = 1 - 2r_{ki}^2$.

related to the amount of actual change in the level of convergence, then it might be hypothesized that, in psychophysical parlance, more JNDs in stability will be perceived in the process of varying the value of r_{ki} from -0.75 to -0.50 than for any other comparable amount of change in r_{ki}. The implications of this fact for the salary administrator would seem to be that costs of varying the amount of perceived stability in $r_{k(k+i)}$ would be greatest for companies who pursue R1 policies, cost being measured in terms of the change required in r_{ki} to produce the effect desired. For example, knowledge that r_{ki} has been equal to 0.70 for a number of years makes any attempt to raise $r_{k(k+i)}$ to a higher level seem uneconomical since increasing the value of r_{ki} will have little effect on $r_{k(k+i)}$ in this range.

The effect of r_{ki} on σ_{k+i} can also be subdivided on the basis of the sign and magnitude of r_{ki}.

(*a*) $r_{ki} \geq 0.0$. Under Policies R1 and R2 the value of σ_{k+i} is a steadily increasing function over time. A brief look at the formula for σ_{k+i}^2 will explain these results. Because σ_i^2 is an additive component, σ_{k+i}^2 will always increase whenever there is variance in the increments. When r_{ki} is zero, the amount of this increase is simply equal to σ_i^2. However, as r_{ki} assumes larger positive values, the increase in the variance of pay will be a function of σ_i^2 plus the covariance term, $2\sigma_i\sigma_k r_{ki}$.

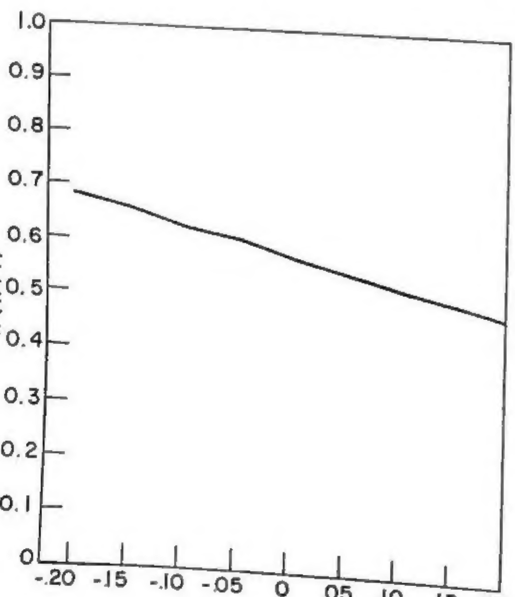


FIG. 33. The mean level of convergence of $r_{k(k+i)}$ averaged across values of r_{ki} plotted as a function of σ_i.

It is clear that the variance of pay will not change if there is no variance in the increments awarded (i.e., $\sigma_i = 0.0$), regardless of the value of r_{ki}. This follows from the mathematical rule that the variance of a distribution remains unchanged if a constant is added or subtracted from each score in the sample.

(*b*) $r_{ki} < 0.0$. Under Policy R3 the level of σ_{k+i} converges instead of steadily increasing or decreasing. The level of convergence is in some cases higher than the initial value of σ_k despite the negative covariance term produced by r_{ki}.

Algebraically, it can be demonstrated that whenever r_{ki} is less than 0.0, and Policy S1 is in effect, the curve of σ_{k+i} becomes asymptotic at a value determined from the following formula: $\sigma_{k+i} = \sigma_i/-2r_{ki}^2$.

Effects of σ_i

Regardless of the policy concerning r_{ki}, it was found that the level of convergence of $r_{k(k+i)}$ decreases as the value of σ_i is progressively enlarged from one year to the next. Figure 33 is a plot of the mean level of convergence for each of the nine levels of the σ_i factor when the results are averaged across all r_{ki} policies. The narrow range of convergence levels of $r_{k(k+i)}$ for the different amounts of change in σ_i indicates that the average amount of stability in relative monetary position does not vary a great deal for different policies controlling σ_i.

As indicated earlier, since σ_i is an additive component of σ_{k+i}, the greater the increase in σ_i, the greater the value of σ_{k+i}. That is, for a given amount of time, regardless of the policy controlling r_{ki}, σ_{k+i} will be larger under an S2 policy than either an S1 or S3 policy.

Effects of σ_i/σ_k

Regardless of the ratio of σ_i to σ_k of the original or starting salaries, the curves of $r_{k(k+i)}$ will eventually converge. That is, σ_i/σ_k does not differentially affect relative monetary status of individuals with the passage of sufficient time. However, the amount of time required by $r_{k(k+i)}$ to achieve stability will vary as a function of σ_i/σ_k. The time necessary to reach the level of $r_{k(k+i)}$ at which convergence

occurs is increased when σ_i/σ_k is increased for $r_{ki} \geq 0.0$, whereas the time is decreased when σ_i/σ_k is increased and $r_{ki} < 0.0$.

As has been noted previously, it is readily apparent from the formula that the value of σ_{k+i} is directly related to the size of σ_i. It is not surprising then that the larger values of σ_i/σ_k were associated with the higher values of σ_{k+i} for all treatment combinations.

These findings indicate that the motivational leverage inherent in σ_i would have to have its effect through broadening the range of expected pay. The effect of σ_i/σ_k on the correlation of salaries on successive years is negligible since the same amount of relative stability is achieved regardless of the value of σ_i.

Discussion

A model of this kind sorts out and clarifies a group of specific variables which turn up under the heading of pay. Unfortunately, pay is complicated; the authors cannot accept any responsibility for the portion of the complication which exists in the real world. Restating the variety of relationships within this complication would seem to help make future research and conceptualizations specific. In addition:

I. The salary administration may want to change either: *(a)* the employee's feeling that he can get ahead (relatively), or *(b)* the size of the jump in the employee's aspiration. The first he does by policies influencing $r_{k,k+i}$; the second by policies influencing σ_{k+i}.

If he wants an employee to feel that he can get ahead (relatively), the administrator remembers:

1. If he manages r_{ki} so that it is ≥ 0.0, $r_{k,k+i}$ will, eventually, be asymptotic to 1.0.
2. Stability of relative position ($r_{k,k+i}$) will be higher, the higher r_{ki}.
3. It will stabilize sooner, the higher r_{ki}.
4. If he operates r_{ki} at 0.0, stability will be largely influenced by σ_i. It will be higher if σ_i is constant or decreasing over time, lower if σ_i increases steadily.
5. The ratio of σ_i/σ_k will have a negligible effect on the level at which $r_{k,k+i}$ stabilizes.
6. Stability in $r_{k,k+i}$ occurs sooner if σ_i/σ_k is increased and $r_{ki} < 0.0$ and later if $r_{ki} \geq 0.0$.

If he wants to broaden aspirations (σ_{k+i}):

1. Obviously, it depends on the size of σ_i.
2. Variation increases faster if $r_{ki} \geq 0.0$.
3. It increases faster, the higher r_{ki}.
4. If $r_{ki} < 0.0$, variation tends to become stable.

The administrator seeking to maximize a feeling of opportunity and high aspirations has a problem. If he lets r_{ki} get high, σ_{k+i} goes up and the employee has wider aspirational horizons. However, $r_{k,k+i}$ tends to rise, and the probability that the employee will advance relative to others diminishes. On the other hand, increasing σ_i systematically will both raise σ_{k+i} and lower $r_{k,k+i}$, presumably what he wanted. He may be comfortable with the assumption that the range of performance (and hence σ_{k+i}) increases, but deliberately lowering r_{ki} and $r_{k,k+i}$ is disturbing. Common sense and experience suggest that repeated measures of performance tend to be correlated. Further, one likes to think that eventually—with experience with people on jobs—pay will predict pay. He has had to relax both these notions. Moreover, as σ_{k+i} increases across levels, it has real implications for the spread of the organization. It is indeed a difficult situation. Perhaps that is why salary administrators' own salaries tend to be high.

II. In the process of providing guide lines for controlling the administration of salaries, the focus has been shifted to those variables which are most appropriately manipulated in order to attain specific objectives. Many salary administrators concern themselves solely with the adjustment of the amount of raises, that is, $\bar{i}$. It is clear that the only thing which is directly affected by this is $\bar{k}$. Many discussions of equity (Adams, 1963; Jaques, 1961) have indicated that the absolute amount of salary is in itself comparatively unimportant without also considering the relative salaries of reference groups. Since the effect of $\bar{i}$ on the more crucial dependent variables in the model is only through its empirically determined correlation with one of the other parameters in each particular situation, it would appear that more effective control of salaries could be achieved by manipulation of σ_i and r_{ki}.

III. In this connection it is suggested that the compensation policy be controlled by

means of adjustments in σ_i rather than r_{ki}. σ_i clearly is a more manageable variable. Whereas one may be able to judge roughly the manner in which raises should be distributed among present salary levels in order to approximate the desired magnitude of r_{ki}, precise manipulation of this parameter is difficult in practice. There is no known algorithm with which to determine the exact method of awarding increments in order to produce a desired level of r_{ki}. However, the results in terms of the level of convergence of $r_{k(k+i)}$ are much alike for a wide range of r_{ki}.

IV. The model has also provided a possible explanation of the conflict which we noted in Chapter 2 between our results and those of Brenner and Lockwood (1965) on the behavior of $r_{k(k+i)}$ with increases in tenure. Brenner and Lockwood found evidence to support the notion that the correlation between salaries over a given amount of time increases with greater tenure, whereas our results indicated that tenure did not affect $r_{k(k+i)}$ in this fashion. Although the conflicting reports may be resolved by attributing the discrepancy in results to differences which existed among the samples and the companies which they represented, systematic investigation of the model has demonstrated the manner in which it is possible for $r_{k(k+i)}$ to steadily increase or decrease over equal units of time. That is, it was noted that when $r_{ki} \geq 0.0$ for all levels of Policies S1 and S3, $r_{k(k+i)}$ increased over each successive unit of time. However, when $r_{ki} < 0.0$, depending upon the values of the other factors investigated, $r_{k(k+i)}$ was found to decrease. The authors are presently reexamining the data from these two studies to test the hypothesis that differences in r_{ki} produced the conflicting results on the test of tenure effects.

Suggestions for Future Research

The authors have indulged in a substantial amount of speculation about the relationship of the constructs embraced by the model of pay to constructs within other realms of psychology. For example, although empirical results were lacking, many tentative conclusions have been based upon the untried assumption that modifications in compensation policies are perceived by the people affected by the changes. The need is quite apparent for a gr deal of additional work on the relationshi the parameters of pay to variables connec with perception, attitudes, and motivation. these future research undertakings, the mo apparently would be useful in operationaliz those aspects or constructs of pay which sho serve as experimental variables.

It should be emphasized, however, that existence of the model will not surmount the obstacles in the path of methodologica sound and persuasive research. Many oth problems which have been mentioned he such as sampling difficulties, the use of termi criteria, and management's chronic relucta to release salary data, will continue to plag researchers concerned with compensation search. Better experimental design and m careful collection of data are problems wh can only be overcome by means of a compl reorientation in the philosophy of resear promoted by the psychologist and adopted those agencies or companies that spons research.

REFERENCES

ADAMS, J. S. Toward an understanding of inequit *Journal of Abnormal and Social Psychology*, 1963, 422–436.

BRENNER, M. H., & LOCKWOOD, H. C. Salary as predictor of salary: A 20-year study. *Journal Applied Psychology*, 1965, 49, 295–298.

FLEISHMAN, E. A., & FRUCHTER, B. Factor structu and predictability of successive stages of learn Morse Code. *Journal of Applied Psychology*, 1960, 97–108.

GUILFORD, J. P. *Psychometric methods*. (2nd ed.) N York: McGraw-Hill, 1954.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. *The motivation to work*. (2nd ed.) New York: Wile 1962.

JAQUES, E. *Equitable payment*. New York: Wiley, 196

LAWLER, E. E., & PORTER, L. W. Perceptions regardin management compensation. *Industrial Relation* 1963, **3**(1), 41–49.

PATTON, A. *Men, money, and motivation*. New York McGraw-Hill, 1961.

PEATMAN, J. G. *Introduction to applied statistics*. Ne York: Harper, 1963.

THORNDIKE, R. *Personnel selection: Test and measure ment techniques*. New York: Wiley, 1949.

VROOM, V. H. *Work and motivation*. New York: Wiley 1964.

WEISSINGER, T. E. Equitable salary structures fo professionals. *Management Record*, 1958, **20**, 162–165

(Received December 24, 1966)